Profesor: Dr. Oldemar Rodríguez Rojas
PF-1319 y PF-1320 Análisis de Datos II
Fecha de Entrega: Domingo 11 de septiembre a las 12 media noche
Instrucciones:

- La solución a cada tarea se debe subir en el aula virtual, no pueden ser enviadas por correo.
- 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.
- Las tareas se pueden entregar tarde, pero cada día de atraso tendrá un rebajo de 20 puntos.

Tarea Número 3

- **Pregunta 1:** [20 puntos] Dada la siguiente Tabla de Testing de un Scoring de Crédito:

MontoCredito	IngresoNeto	CoefCreditoAvaluo	MontoCuota	GradoAcademico	BuenPagador	Predicción KNN
17341	1	11	Alto	Licenciatura	Si	Si
18315	1	11	Alto	Licenciatura	Si	No
19172	1	11	Alto	Licenciatura	Si	Si
16761	1	11	Alto	Licenciatura	Si	Si
274224	1	1	Alto	Licenciatura	Si	Si
15998	1	11	Alto	Licenciatura	Si	Si
11669	1	11	Alto	Licenciatura	Si	Si
19875	2	11	Alto	Licenciatura	Si	Si
19381	1	11	Alto	Licenciatura	Si	No
26934	1	11	Alto	Licenciatura	Si	Si
13733	2	12	Alto	Licenciatura	Si	Si
38185	2	11	Alto	Licenciatura	Si	Si
40644	2	12	Alto	Licenciatura	Si	Si
13316	2	12	Alto	Bachiller	Si	Si
16469	1	12	Alto	Bachiller	Si	Si
15747	2	12	Alto	Bachiller	Si	Si
16331	1	12	Alto	Bachiller	Si	Si
27503	2	12	Alto	Bachiller	Si	Si
12873	2	12	Alto	Bachiller	Si	Si
63316	2	11	Medio	Bachiller	No	No
123772	2	11	Alto	Bachiller	No	No
21449	2	11	Alto	Bachiller	No	Si
12325	1	11	Bajo	Bachiller	No	Si
16229	2	6	Bajo	Bachiller	No	No
40170	1	12	Medio	Licenciatura	No	No

1. Usando la columna `BuenPagador` en donde aparece el verdadero valor de la variable a predecir y la columna `Predicción KNN` en donde aparece la predicción del Método `KNN` para esta tabla de Testing, calcule la Matriz de Confusión.
2. Con la Matriz de Confusión anterior calcule "a mano" la Precisión Global, el Error Global, la Precisión Positiva (PP), la Precisión Negativa (PN), la Proporción de Falsos Positivos (PFP), la Proporción de Falsos Negativos (PFN), la Asertividad Positiva (AP) y la Asertividad Negativa (AN).

- **Pregunta 2:** [20 puntos] Programe en lenguaje **Python** una clase que contenga un método que reciba como entrada la Matriz de Confusión (para el caso 2×2) que calcule y retorne en un diccionario: la Precisión Global, el Error Global, la Precisión Positiva (PP), la Precisión Negativa (PN), la Proporción de Falsos Positivos (PFP), la Proporción de Falsos Negativos (PFN), la Asertividad Positiva (AP) y la Asertividad Negativa (AN).

Supongamos que tenemos un modelo predictivo para detectar Fraude en Tarjetas de Crédito, la variable a predecir es `Fraude` con dos posibles valores `Sí` (para el caso en que sí fue fraude) y `No` (para el caso en que no fue fraude). Supongamos que la matriz de confusión es:

	`No`	`Sí`
`No`	892254	252
`Sí`	9993	270

- Con ayuda de la clase programada anteriormente calcule la Precisión Global, el Error Global, la Precisión Positiva (PP), la Precisión Negativa (PN), la Proporción de Falsos Positivos (PFP), la Proporción de Falsos Negativos (PFN), la Asertividad Positiva (AP) y la Asertividad Negativa (AN).
- ¿Es bueno o malo el modelo predictivo? Justifique su respuesta.

- **Pregunta 3:** [20 puntos] En este ejercicio usaremos la tabla de datos `abandono_clientes.csv`, que contiene los detalles de los clientes de un banco.

  La tabla contiene 11 columnas (variables), las cuales se explican a continuación.

  - `CreditScore:` Indica el puntaje de crédito.
  - `Geography:` País al que pertenece.
  - `Gender:` Género del empleado.
  - `Age:` Edad del empleado.
  - `Tenure:` El tiempo del vínculo con la empresa.
  - `Balance:` La cantidad que les queda.
  - `NumOfProducts:` Los productos que posee.
  - `HasCrCard:` Tienen tarjeta de crédito o no.
  - `IsActiveMember:` Es un miembro activo o no.
  - `EstimatedSalary:` Salario estimado.
  - `Exited:` Indica si el cliente se queda o se va.

  Realice lo siguiente:

  1. Cargue en **Python** la tabla de datos `abandono_clientes.csv`.
  2. ¿Es este problema equilibrado o desequilibrado? Justifique su respuesta.
  3. Use el método de K vecinos más cercanos en **Python** para generar un modelo predictivo para la tabla `abandono_clientes.csv` usando el 75 % de los datos para la tabla aprendizaje y un 25 % para la tabla testing. Intente con varios valores de K e indique cuál fue la mejor opción.
  4. Genere un Modelo Predictivo usando K vecinos más cercanos para cada uno de los siguientes núcleos `ball_tree, kd_tree` y `brute` ¿Cuál produce los mejores resultados en el sentido de que predice mejor?

- **Pregunta 4:** [20 puntos] En este ejercicio vamos a usar la tabla de datos `raisin.csv`, que contiene es resultado de un sistema de visión artificial para distinguir entre dos variedades diferentes de pasas (Kecimen y Besni) cultivadas en Turquía. Estas imágenes se sometieron a varios pasos de preprocesamiento y se realizaron 7 operaciones de extracción de características morfológicas utilizando técnicas de procesamiento de imágenes.

  El conjunto de datos tiene 900 filas y 8 columnas las cuales se explican a continuación.

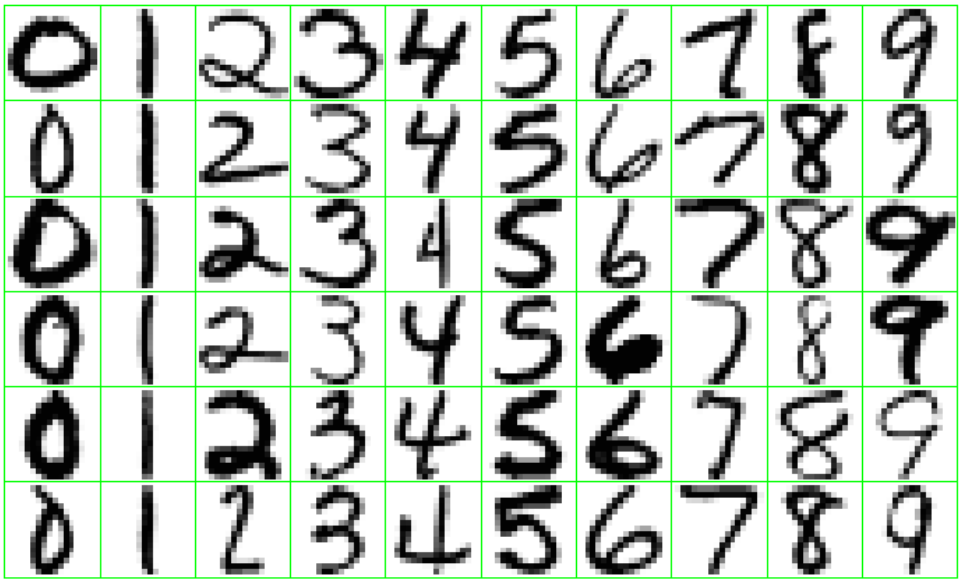
  - `Area`: El número de píxeles dentro de los límites de la pasa..
  - `MajorAxisLength`: La longitud del eje principal, que es la línea más larga que se puede dibujar en la pasa.
  - `MinorAxisLength`: La longitud del eje pequeño, que es la línea más corta que se puede dibujar en la pasa.
  - `Eccentricityl`: Una medida de la excentricidad de la elipse, que tiene los mismos momentos que las pasas.
  - `ConvexArea`: El número de píxeles de la capa convexa más pequeña de la región formada por la pasa.
  - `Extent`: La proporción de la región formada por la pasa al total de píxeles en el cuadro delimitador.
  - `Perimeter`: Mide el entorno calculando la distancia entre los límites de la pasa y los píxeles que la rodean.
  - `Class`: Tipo de pasa Kecimen y Besni (Variable a predecir).

  Realice lo siguiente:

  1. Cargue en **Python** la tabla de datos `raisin.csv`.
  2. Realice un análisis exploratorio (estadísticas básicas) que incluya: el resumen numérico (media, desviación estándar, etc.), la correlación entre las variables, el poder predictivo de las variables predictoras. Interprete los resultados.
  3. ¿Es este problema equilibrado o desequilibrado? Justifique su respuesta.
  4. Use el método de K vecinos más cercanos en **Python** (con los parámetros por defecto) para generar un modelo predictivo para la tabla `raisin.csv` usando el 75 % de los datos para la tabla aprendizaje y un 25 % para la tabla testing, luego calcule para los datos de testing la matriz de confusión, la precisión global y la precisión para cada una de las dos categorías. ¿Son buenos los resultados? Explique.
  5. Repita el item *4)*, pero esta vez, seleccione las 4 variables que, según su criterio, tienen mejor poder predictivo.
  6. Usando la función programada en el ejercicio 1 y los modelos generados arriba, construya un `DataFrame` de manera que en cada una de las filas aparezca un modelo predictivo y en las columnas aparezcan los índices *Precisión Global, Error Global, Precisión Positiva (PP), Precisión Negativa (PN), Falsos Positivos (FP), los Falsos Negativos (FN), la Asertividad Positiva (AP) y la Asertividad Negativa (AN)*. ¿Cuál de los modelos es mejor para estos datos?

7. Repita el item *4)*, pero esta vez en el método `KNeighborsClassifier` utilice los 3 diferentes algoritmos `ball_tree, kd_tree` y `brute`. ¿Cuál da mejores resultados?

- **Pregunta 5:** [20 puntos] En este ejercicio vamos a predecir números escritos a mano (Hand Written Digit Recognition), la tabla de aprendizaje está en el archivo `ZipDataTrainCod.csv` y la tabla de testing está en el archivo `ZipDataTestCod.csv`. En la figura siguiente se ilustran los datos:



Los datos de este ejemplo vienen de los códigos postales escritos a mano en sobres del correo postal de EE.UU. Las imágenes son de 16×16 en escala de grises, cada pixel va de intensidad de -1 a 1 (de blanco a negro). Las imágenes se han normalizado para tener aproximadamente el mismo tamaño y orientación. La tarea consiste en predecir, a partir de la matriz de 16×16 de intensidades de cada pixel, la identidad de cada imagen $(0, 1, \ldots, 9)$ de forma rápida y precisa. Si es lo suficientemente precisa, el algoritmo resultante se utiliza como parte de un procedimiento de selección automática para sobres. Este es un problema de clasificación para el cual la tasa de error debe mantenerse muy baja para evitar la mala dirección de correo. La columna 1 tiene la variable a predecir `Número` codificada como sigue: 0=‘cero’; 1=‘uno’; 2=‘dos’; 3=‘tres’; 4=‘cuatro’; 5=‘cinco’;6=‘seis’; 7=‘siete’; 8=‘ocho’ y 9=‘nueve’, las demás columnas son las variables predictivas, además cada fila de la tabla representa un bloque 16×16 por lo que la matriz tiene 256 variables predictoras.

1. Usando K vecinos más cercanos genere un modelo predictivo para la tabla de aprendizaje, con los parámetros que usted estime más convenientes.
2. Con la tabla de testing calcule la matriz de confusión, la precisión global, el error global y la precisión en cada unos de los dígitos. ¿Son buenos los resultados?
3. Repita los items *1)* y *2)* pero usando solamente los 1s, 6s y los 9s. ¿Mejora la predicción?
4. Repita los items *1)* y *2)* utilizando `n_neighbors=5` y `algorithm=‘‘auto’’` (parámetros por defecto) pero reemplazando cada bloque 4×4 de píxeles por su promedio. ¿Mejora la predicción? Recuerde que cada bloque 16×16 está representado por una fila en las

matrices de aprendizaje y testing. **Despliegue la matriz de confusión resultante**. La matriz de confusión obtenida debería ser:

	cero	uno	dos	tres	cuatro	cinco	seis	siete	ocho	nueve
cero	343	6	2	0	0	0	3	1	4	0
uno	1	250	1	0	5	1	1	3	2	0
dos	5	3	180	1	0	1	0	1	6	1
tres	2	1	3	138	0	15	0	1	4	2
cuatro	0	4	4	0	166	1	2	0	1	22
cinco	10	0	1	17	0	123	0	1	5	3
seis	6	1	2	0	2	2	157	0	0	0
siete	0	0	2	0	7	0	0	130	1	7
ocho	9	19	2	2	1	5	2	0	124	2
nueve	1	0	0	0	2	0	0	1	1	172

No es necesario que las categorías se muestren en orden.

5. Repita los items *1)* y *2)* pero reemplazando cada bloque $p \times p$ de píxeles por su promedio. ¿Mejora la predicción? (pruebe con algunos valores de p). **Despliegue las matrices de confusión resultantes**.

oldemar rodríguez
CONSULTOR en M1NƐR14 DƐ D4T0S