Profesor: Dr. Oldemar Rodríguez Rojas
PF-1319 y PF-1320 Análisis de Datos II
Fecha de Entrega: Domingo 9 de octubre 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 7

1. **[30 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) Use Bayes en Python para generar un modelo predictivo para la tabla `novatosNBA.csv` usando el 80 % de los datos para la tabla aprendizaje y un 20 % para la tabla testing. Obtenga los índices de precisión e interprete los resultados.

   b) Construya un DataFrame que compare el modelo generado en el ítem anterior contra los modelos vistos en las clases anteriores para la tabla `novatosNBA.csv`. Para esto en cada una de las filas debe aparecer un modelo predictivo y en las columnas aparezcan los índices *Precisión Global*, *Error Global*, *Precisión Positiva (PP)* y *Precisión Negativa (PN)*. ¿Cuál de los modelos es mejor para estos datos?

2. **[30 puntos]** Este conjunto de datos es originalmente del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales. El objetivo del conjunto de datos es predecir de forma diagnóstica si un paciente tiene diabetes o no, basándose en determinadas medidas de diagnóstico incluidas en el conjunto de datos. El conjunto de datos tiene 390 filas y 16 columnas:

   - **X:** Id del paciente.
   - **colesterol:** Colesterol en mg/dL.
   - **glucosa:** Glucosa en mg/dL.
   - **hdl_col:** Lipoproteínas (colesterol bueno).
   - **prop_col_hdl:** Proporción del colesterol entre el hdl.
   - **edad:** Edad del paciente.
   - **genero:** Género del paciente.
   - **altura:** Altura en pulgadas del paciente.
   - **peso:** Peso en libras del paciente.

- **IMC:** índice de masa corporal.
- **ps_sistolica:** Presión arterial sistólica.
- **ps_diastolica:** Presión arterial diastólica.
- **cintura:** Longitud de la cintura en pulgadas.
- **cadera:** Longitud de la cadera en pulgadas.
- **prop_cin_cad:** Proporción de la longitud de la cintura entre la longitud de la cadera.
- **diabetes:** Diagnóstico de la diabetes.

Realice lo siguiente:

a) Cargue en Python la tabla de datos `diabetes.csv`.

b) Use Bayes en `Python` para generar un modelo predictivo para la tabla `diabetes.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.

c) Construya un DataFrame que compare el modelo generado en el ítem anterior contra los modelos vistos vistos en las clases anteriores para la tabla `diabetes.csv`. Para esto en cada una de las filas debe aparecer un modelo predictivo y en las columnas aparezcan los índices *Precisión Global*, *Error Global*, *Precisión Positiva (PP)* y *Precisión Negativa (PN)*. ¿Cuál de los modelos es mejor para estos datos?

d) Repita el ítem 2, pero esta vez seleccione 6 variables predictoras ¿Mejora la predicción?

3. [**20 puntos**] Para la siguiente tabla, la cual se vio en clase, suponga que se tiene una nueva fila o registro de la base de datos `t = (Isabel, F, 4, ?)`, prediga (a mano) si Isabel corresponde a la clase pequeño, mediano o alto.

Nombre	Género	Altura	Clase
Kristina	F	1	P
Jim	M	5	A
Maggi	F	4	M
Martha	F	4	M
Stephanie	F	2	P
Bob	M	4	M
Kathy	F	1	P
Dave	M	2	P
Worth	M	6	A
Steven	M	6	A
Debbie	F	3	M
Todd	M	5	M
Kim	F	5	M
Amy	F	3	M
Wynette	F	3	M

Id	Monto.Crédito	Ingreso.Neto	Monto.Cuota	Grado.Académico	Buen.Pagador
1	2	4	1	4	Sí
2	2	3	1	4	Sí
3	4	1	4	2	No
4	1	4	1	4	Sí
5	3	3	3	2	No
6	3	4	1	4	Sí
7	4	2	3	2	No
8	4	1	3	2	No
9	3	4	1	3	Sí
10	1	3	2	4	Sí
11	1	4	2	4	Sí

4. [**20 puntos**] Para la siguiente tabla, la cual se vio en clase, suponga que se tiene una nueva fila o registro `12 = (1, 3, 2, 4, ?)` en la base de datos, prediga (a mano) si el individuo corresponde a un buen pagador o a un mal pagador.