EXAMEN FINAL DE MINERIA DE DATOS

1.Clustering.

Utilizando el dataset “Wholesale customers dataset.csv”

El conjunto de datos se refiere a clientes de un distribuidor mayorista. Incluye el gasto anual en unidades monetarias (mu) en diversas categorías de productos.

Attribute Information:

1) FRESH: annual spending (m.u.) on fresh products (Continuous);  
2) MILK: annual spending (m.u.) on milk products (Continuous);  
3) GROCERY: annual spending (m.u.)on grocery products (Continuous);  
4) FROZEN: annual spending (m.u.)on frozen products (Continuous)  
5) DETERGENTS\_PAPER: annual spending (m.u.) on detergents and paper products (Continuous)  
6) DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous);  
7) CHANNEL: customersâ€™ Channel - Horeca (Hotel/Restaurant/CafÃ©) or Retail channel (Nominal)  
8) REGION: customersâ€™ Region â€“ Lisnon, Oporto or Other (Nominal)

1. Elimine las columnas channel y región.
2. Revise si hay nulos
3. Grafique los histogramas de las columnas.
4. Determine la cantidad optima de clúster con el dendrograma.
5. Determine la cantidad de clúster con el método del codo.
6. Escale los datos
7. Construya el modelo utilizando Clúster jerárquico
8. Cree un dataframe llamado resultados, Agregando una columna llamada jerárquico al dataframe resultados, con la predicción del modelo jerárquico
9. Construya otro modelo utilizando k-means
10. Agregue una columna al dataframe resultados con una columna que contenga la predicción del modelo k-means, llame la columna k-means.

2. Redes Neuronales Artificiales

Se va a utilizar el conjunto de datos sobre el inicio de la diabetes en los indios Pima. Describe los datos de los registros médicos de los pacientes de los indios Pima y si tenían la aparición de la diabetes en un plazo de cinco años. Se trata de un problema de clasificación binaria (aparición de diabetes como 1 o no como 0.

Las variables de entrada que describen a cada paciente son numéricas y tienen variables de escalas. A continuación se enumeran los ocho atributos del set de datos:

1. Número de embarazos.  
2. Concentración de glucosa en plasma a 2 horas en una prueba oral de tolerancia a la glucosa.  
3. Presión arterial diastólica (mm Hg).  
4. Grosor del pliegue de la piel del tríceps (mm).  
5. Insulina sérica de 2 horas (mu U/ml).  
6. Índice de masa corporal.  
7. Función del pedigrí de la diabetes.  
8. Edad (años).  
9. Clase, aparición de diabetes en un plazo de cinco años.

a) X son las 8 primeras columnas e Y la última columna

b) Cree un RNA con tres capas. Las dos primeras con RELU y la última con Sigmoide como funciones de activación. La primera capa tiene 8 variables de entrada.

c) Compile el modelo utilizando estos parámetros: loss='binary\_crossentropy', optimizer='adam', metrics=['accuracy']

d) ajuste el modelo con estos parámetros : epochs=150, batch\_size=10

e) evalue el modelo utilizando la siguiente instrucción.

scores = model.evaluate(X, Y)

print("\n%s: %.2f%%" % (model.metrics\_names[1], scores[1]\*100))

f) calcule el y\_predict

g) encuentre la matriz de confusión.