UNIVERSIDAD DE COSTA RICA

SISTEMAS DE ESTUDIO DE

POSTGRADO EN ESTADÍSTICA

Examen parcial n° 2

Este debe ser entregado en formato .PDF, además de entregar el archivo .RMD, el enlace del Data Studio + .PDF de los dashboard en Data Studio.

Nombre: _____ Carné: _____

I. RESPUESTAS BREVES. (50 PUNTOS / 5 PTS CU).

1. ¿En términos de unión de elementos en un cluster jerárquico, cuál es la diferencia de utilizar un método Complete o Maximun y un método single o mínimum? Explíquelo si así lo requiere con una figura o imagen, o con notación matemática la forma de proceder. (5pts)
2. Explique en términos de distancia, entre un método jerárquico y uno de reasignación, en que consiste la formación de clúster en cada método, en que se asemejan y cuáles son sus diferencias. (5pts)
3. Si tuviera que utilizar un método de agrupamiento que por análisis se sabe que siempre posee entre 5 – 7 densidades en el área en un plano 2D, ¿cuál sería mejor método? Justifique su respuesta. (5pts)
4. Si quisiera analizar un conjunto de datos que posee tanto variables cualitativas como cuantitativas, ¿cuál sería el método de distancia ideal? Exponga la fórmula matemática explicando sus elementos. (5pts)
5. Explique en que consiste la tokenización, el corpus y la matriz de términos, luego explique por qué es que es preferible utilizar el corpus en el análisis del Text Mining.
6. ¿Podríamos realizar un análisis de sentimientos con solo ya sea sentimientos negativos o solo positivos? Dependiendo de su respuesta, que haría para solo analizar un tipo de sentimiento (positivo o negativo).
7. En el análisis de sentimiento, qué es el lexicón, que quiere decir la asignación numérica a una palabra, y cree que se podría cambiar el lexicón este análisis.
8. Explique cuál es la tipología de un árbol. Cuáles son los nodos que determinan así el espacio de asignación de los casos a un cierto valor promedio o valor de clasificación.

9. Explique por qué en un árbol de decisión, se puede utilizar para las variables independientes tanto métricas cuantitativas y cualitativas, sin que esto tenga un efecto significativo en la estimación del modelo.
10. Explique qué es el método de validación de Visual Assessment of cluster Tendency, y en qué consiste, y por qué nos permite decir que es pertinente realizar un análisis de cluster.

II. PREGUNTAS PRÁCTICAS[1] (50 pts / 10 PTS CU).

1. A partir del archivo “Medicina empresarial --- IAM”, se le pide que realice un Dashboard en Data Studio, bajo las siguientes indicaciones (10 PTS).

    a. Se quiere resaltar el monto promedio de la consulta y de las incapacidades promedios en días, estos como datos que deben sobresalir.
    b. Se quiere estadísticas descriptivas para conocer la información demográfica de la población (edad, sexo, tipo de trabajador).
    c. Interesa conocer el área de trabajo, y en esta saber ante todo por área cuál es el % de grasa, como se constituye por área las citas, y la condición % de grasa
    d. Se quiere conocer de forma general, para el mes de junio, cuáles fueron los primeros 5 padecimientos crónicos, los primeros 5 diagnósticos micro y macro.
    e. Finalmente se quiere conocer la relación entre la instancia de atención, y si el cliente fue entonces referido a Laboratorio, rayos “X” o consulta adicional.
    f. Interesa tomar como segmentadores, asociadas a todos los resultados, el tipo de trabajador, el sexo, el área de trabajo.
    g. Haga una conclusión de no más de 10 líneas sobre la situación para el mes de junio.

2. A partir del archivo “protein”, se le pide que realice un análisis de cluster bajo las siguientes indicaciones (10 PTS).

    a. Realice las estadísticas descriptivas de las variables cuantitativas.
    b. Realce una de las pruebas para saber si es pertinente hacer un análisis de cluster
    c. Realice el cluster por k-medias. ¿Cuántos grupos deberíamos tener?
    d. Realice el cluster jerárquico. ¿Cuántos grupos debemos tener?
    e. Valido los resultados
    f. Perfile los resultados de la segmentación.

[1] No olvide explicar muy brevemente cada respuesta.

g. Realice un resumen no mayor de 5 líneas donde exponga lo principal del análisis.

3. A partir del discurso de Carlos Alvarado en la XXVI Cumbre Iberoamericana (archivo .txt), se le pide que realice un análisis descriptivo de text mining descriptivo bajo las siguientes indicaciones (10 PTS).

    a. Realice la limpieza de datos.
    b. Tokenizar los datos y realizar una distribución de frecuencias de las palabras.
    c. Meter los datos en un Corpus y luego en una matriz de términos.
    d. Realice dos nueves de palabras: una con todas las palabras, y otra eliminando todas las palabras como artículos definidos, indefinidos, determinaciones, etc.
    e. Realice un análisis de sentimiento utilizando el léxico_afinn.en.es. Determine los principales sentimientos negativos como positivos del discurso.
    f. Realice un resumen no mayor de 5 líneas donde exponga los principales resultados del discurso de Carlos Alvarado en la XXVI Cumbre Iberoamericana.

4. A partir del archivo “credit”, realice el siguiente árbol de clasificación bajo las siguientes indicaciones (10 PTS).

    a. Tome como árbol uno con variable dependiente “credit_history” y como independientes “purpose”, “employment_duration”, y “housing”.
    b. Realice las estradísticas descriptivas de las variables puestas en causa.
    c. Parte el archivo de datos en 75% entrenamiento y 25% validación
    d. Estime el árbol de decisión de clasificación y su representación gráfica. Determine e interprete los 2 nodos terminales que considere más importantes.
    e. Evalúe el modelo con tablas de confusión para la parte de entrenamiento y para la parte de validación.
    f. Realice por lo menos algún cambio de ciertos hyper-parámetros y verifique si el modelo mejoró.
    g. Realice un resumen no mayor de 5 líneas donde exponga los principales resultados del árbol de clasificación.

5. A partir del archivo “credit”, realice el siguiente árbol de regresión bajo las siguientes indicaciones (10 PTS).

    a. Tome como árbol uno con variable dependiente “amount” y como independientes “months_loan_duration”, “years_at_residence”, “age” y “housing”
    b. Realice las estradísticas descriptivas de las variables puestas en causa.
    c. Parte el archivo de datos en 60% entrenamiento y 40% validación.
    d. Estime el árbol de regresión de y su representación gráfica. Determine e interprete los 3 nodos terminales que considere más importantes.
    e. Evalúe el modelo con tablas de confusión para la parte de entrenamiento y para la parte de validación.
    f. Realice por lo menos algún cambio de ciertos hyper-parámetros y verifique si el modelo mejoró.
    g. Realice un resumen no mayor de 5 líneas donde exponga los principales resultados del árbol de clasificación.

¡BUENA SUERTE!