CUESTIONARIO 2

GRUPO # 2

PRUEBAS DE CONCEPTO

1. ¿Qué es Clustering?

El **Clustering** es una tarea que tiene como finalidad principal lograr el agrupamiento de conjuntos de objetos no etiquetados, para lograr construir subconjuntos de datos conocidos como Clusters

2. ¿Cómo trabaja el algoritmo de clústeres de Microsoft?

El algoritmo de clústeres de secuencia de Microsoft es un algoritmo único que combina el análisis secuencial con la agrupación en clústeres. Puede usar este algoritmo para explorar datos que contienen eventos que pueden vincularse con rutas o *secuencias*. El algoritmo encuentra las secuencias más comunes y realiza una agrupación en clústeres para buscar secuencias que sean similares.

3. ¿Cuáles son los datos requeridos para los modelos de agrupación en clústeres?

Los requisitos para un modelo de agrupación en clústeres son los siguientes:

- **Una columna de una sola clave:** cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.
- **Columnas de entrada**: Cada modelo debe tener al menos una columna de entrada que contenga los valores que se utilizan para generar los clústeres. Puede tener tantas columnas de entrada como desee, pero dependiendo del número de valores existentes en cada columna, la adición de columnas adicionales podría aumentar el tiempo necesario para entrenar el modelo.
- **Una columna de predicción opcional:** El algoritmo no necesita una columna de predicción para generar el modelo, pero puede agregar una columna de predicción de casi cualquier tipo de datos. Los valores de la columna de predicción se pueden tratar como entradas del modelo de agrupación en clústeres, o se puede especificar que solo se utilicen para las predicciones.

4. ¿Qué ventajas existen en la implementación del algoritmo de clústeres?

- El número óptimo de clústeres se puede obtener por el mismo modelo, a través de su muy útil dendrograma.
- Fácil de entender, fácil de adaptar. Trabaja bien con conjuntos de datos grandes o pequeños, es eficiente y tiene buen desempeño.

5. ¿Qué desventajas existen en la implementación del algoritmo de clústeres?

- No es conveniente para grandes conjuntos de datos.
- Necesita que el usuario defina el número de clústeres.

6. ¿Qué es el algoritmo de Arboles de decisión?

Es un modelo predictivo que mapea observaciones sobre un artículo a conclusiones sobre el valor objetivo del artículo. Es uno de los enfoques de modelado predictivo utilizadas en estadísticas, minería de datos y aprendizaje automático

7. ¿Qué hace el Algoritmo de Arboles de decisión para los atributos discretos?

Para los atributos discretos, el algoritmo hace predicciones basándose en las relaciones entre las columnas de entrada de un conjunto de datos.

Específicamente, el algoritmo identifica las columnas de entrada que se correlacionan con la columna de predicción.

8. ¿Cómo funciona el Algoritmo de árbol de decisión?

En pocas palabras, un árbol de decisión es un árbol en el que cada nodo de rama representa una elección entre varias alternativas y cada nodo de hoja representa una decisión.

Los árboles de decisión aprenden y se entrenan a partir de ejemplos dados y predicen para circunstancias no vistas

9. ¿Cuáles son los requisitos para un modelo de árbol de decisión?

Son:

Una columna de una sola clave: cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.

Una columna de predicción: Se requiere al menos una columna de predicción. Puede incluir varios atributos de predicción en un modelo y pueden ser de tipos diferentes, numérico o discreto. Sin embargo, el incremento del número de atributos de predicción puede aumentar el tiempo de procesamiento.

Columnas de entrada: Se requieren columnas de entrada, que pueden ser discretas o continuas. Aumentar el número de atributos de entrada afecta al tiempo de procesamiento.

10. ¿Ventajas y Desventajas de árbol de decisión?

Ventajas:

- Plantean el problema para que todas las opciones sean analizadas.
- Permiten analizar totalmente las posibles consecuencias de tomar una decisión.
- Proveen un esquema para cuantificar el costo de un resultado y la probabilidad de que suceda.
- Ayuda a realizar las mejores decisiones sobre la base de la información existente y de las mejores suposiciones.
- Provee una estructura sumamente efectiva dentro de la cual se puede estimar cuales son las opciones e investigar las posibles consecuencias de seleccionar cada una de ellas
- Nos ayuda a realizar las mejores decisiones sobre la base de la información existente y de las mejores suposiciones

Desventajas:

- Sólo es recomendable para cuando el número de acciones es pequeño y no son posibles todas las combinaciones.
- En la elección de un modelo, existe una cantidad muy limitada y dificulta para elegir el árbol óptimo.
- Presenta inconvenientes cuando la cantidad de alternativas es grande y cuanto las decisiones no son racionales.
- Al no tener claridad de objetivos, es difícil de organizar las ideas.

11. ¿Qué es Naive Bayes?

En teoría de la probabilidad y minería de datos, un clasificador Naive Bayes es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis simplificadoras adicionales.

Es a causa de estas simplificaciones, que se suelen resumir en la hipótesis de independencia entre las variables predictoras, que recibe el apelativo de *naive*, es decir, *ingenuo*.

12. ¿Ventajas de Naive Bayes?

- Es fácil y rápido predecir la clase de conjunto de datos de prueba. También funciona bien en la predicción multiclase.
- Cuando se mantiene la suposición de independencia, un clasificador Naive Bayes funciona mejor en comparación con otros modelos como la Regresión Logística y se necesitan menos datos de entrenamiento.
- Funciona bien en el caso de variables de entrada categóricas comparada con variables numéricas.

13. ¿Desventajas de Naive Bayes?

- Si la variable categórica tiene una categoría en el conjunto de datos de prueba, que no se observó en el conjunto de datos de entrenamiento, el modelo asignará una probabilidad de 0 y no podrá hacer una predicción. Esto se conoce a menudo

como frecuencia cero. Para resolver esto, podemos utilizar la técnica de alisamiento.

- Otra limitación de Naive Bayes es la asunción de predictores independientes. En la vida real, es casi imposible que obtengamos un conjunto de predictores que sean completamente independientes.

14. ¿Qué son los modelos Naive Bayes?

Estos modelos son llamados algoritmos "Naive", o "Inocentes" en español. En ellos se asume que las variables predictoras son independientes entre sí. En otras palabras, que la presencia de una cierta característica en un conjunto de datos no está en absoluto relacionada con la presencia de cualquier otra característica.

15. ¿Cómo se consiguen los modelos Naive Bayes?

Lo consiguen proporcionando una forma de calcular la probabilidad 'posterior' de que ocurra un cierto evento *A,* dadas algunas probabilidades de eventos 'anteriores'.

$$P(A|R) = \frac{P(R|A)P(A)}{P(R)}$$

- P(A): Probabilidad de A
- P(R|A): Probabilidad de que se de R dado A
- P(R): Probabilidad de R
- P(A|R): Probabilidad posterior de que se de A dado R

16. ¿Qué es el algoritmo de regresión lineal?

La regresión lineal es un algoritmo de aprendizaje supervisado que se utiliza en Machine Learning y en estadística. En su versión más sencilla, lo que haremos es "dibujar una recta" que *nos indicará la tendencia* de un conjunto de datos continuos (si fueran discretos, utilizaríamos Regresión Logística).

17. ¿De qué tipo es el algoritmo de regresión lineal?

El algoritmo de Regresión Lineal es de tipo de aprendizaje supervisado.

18. ¿Qué criterios tiene que cumplir las variables?

Los criterios son 4 y son:

- Las variable de entrada y salida no son ruidosa(se recomienda hacer un método de limpieza de datos antes de aplicar este algoritmo)
- Debe existir una relación lineal entre las variables dependiente e independiente.
- El modelo hará predicciones más confiables si sus variables tienen una distribución normal.
- El modelo se ajustará a los datos cuando tenga variables de entrada altamente correlacionadas

19. ¿En cuántas partes se divide o clasifica el algoritmo de regresión lineal?

Existen diferentes tipos de regresión lineal que se clasifican de acuerdo a sus parámetros:

Regresión lineal simple

- Este modelo sólo está conformado por dos variables estadísticas llamadas X y Y . Para la regresión lineal simple, se asume que X y Y se relacionan mediante la relación funcional

Estimación de los parámetros

- Dado que los parámetros son constantes desconocidas, estas deben estimarse mediante los datos de la muestra, supóngase que se tiene datos , se estimarán los parámetros utilizando el método de mínimos cuadrados.

20. ¿Cuáles son los requisitos para aplicar este algoritmo?

Los requisitos son los siguientes:

- Una única columna key. - Cada modelo debe contener una columna numérica o de texto que identifique cada registro de manera única. No están permitidas las claves compuestas.
- Una columna de predicción. - Se requiere al menos una columna de predicción. Se pueden incluir varios atributos de predicción en un modelo, pero deben ser tipos de datos numéricos continuos. No se puede utilizar un tipo de datos de fecha y hora como atributo de predicción, aunque el almacenamiento nativo para los datos sea numérico.

_____

BERTHY VARGAS VILLARREAL