Análisis de datos con Python

Reactivos

Pregunta 1

El poder identificar con cierto grado de certeza lo que sucederá, involucra cierto conocimiento estadístico y de probabilidades, y puede decirse que se refiere a:

Opciones:

a. Estimación.
b. Pronóstico.
c. Tendencia.
d. Variabilidad.

Pregunta 2

La razón principal del porqué los sistemas de información requieren el manejo de los datos en formatos diversos para su almacenamiento y procesamiento es:

Opciones:

a. Los datos siempre cambian de tamaño y los campos deben ajustarse para recibir la nueva información.
b. El usuario puede decidir utilizar valores de texto para representar números y esto implica cambiar de formato al momento.
c. Existen diferentes tipos de ellos, lo que genera restricciones de espacio en la memoria.
d. Con un tipo de dato de gran capacidad, se puede almacenar desde un valor pequeño, hasta archivos adjuntos en diferentes formatos.

Pregunta 3

Un tipo de dato cualitativo se utiliza para representar características o atributos que no pueden ser cuantificados numéricamente, sino que se dividen en categorías o clases. ¿Cuál de las siguientes es un ejemplo de un tipo de dato cualitativo?

Opciones:

a. Edad de los estudiantes en una clase.
b. Temperatura en grados Celsius.
c. Número de teléfono de una persona.
d. Color de los ojos de una persona.

Pregunta 4

¿Cuál de las siguientes afirmaciones sobre la desviación estándar piensas que es acertada?

Opciones:

a. La desviación estándar siempre es igual al valor máximo del conjunto de datos.
b. La desviación estándar mide la dispersión de un conjunto de datos y representa la raíz cuadrada de la varianza.
c. La desviación estándar es igual a la media de un conjunto de datos.
d. La desviación estándar mide la tendencia central de un conjunto de datos.

Pregunta 5

Un conjunto de datos que representa las edades de un grupo de estudiantes de una escuela secundaria muestra un valor extremadamente alto, por ejemplo, una edad de 99 años, mientras que la mayoría de las edades oscilan entre 12 y 18 años.

Opciones:

a. Un valor que está muy por encima o por debajo de la mayoría de las edades en el conjunto de datos.
b. Un valor que representa la mediana de las edades del grupo de estudiantes.
c. Un valor que representa una edad dentro del rango de 12 a 18 años.
d. Un valor que representa la edad promedio del grupo de estudiantes.

Pregunta 6

Durante el proceso de visualización de datos, se dice que los científicos de datos combinan principalmente estos tres componentes:

Opciones:

a. Historia, tabulares y consolidados.
b. Datos, historia y tabulares.
c. Tabulares, consolidados y resúmenes
d. Datos, visuales e historia.

Pregunta 7

Dentro de las herramientas utilizadas para la distribución de datos, se encuentran los histogramas, que se pueden definir como:

Opciones:

a. Una tabla que lista los datos en orden alfabético.
b. Un gráfico que representa la distribución de frecuencia de datos numéricos agrupados en intervalos.
c. Una representación gráfica de datos que muestra la relación entre dos variables.
d. Un gráfico de barras que compara diferentes categorías de datos.

Pregunta 8

Como ayuda, dentro del lenguaje Python se cuenta con una librería que ayuda para visualizar datos en forma gráfica, como puede ser por medio de un histograma:

Opciones:

a. Spacy
b. PyTorch
c. Seaborn
d. Flask

Pregunta 9

¿Qué son los outliers en el análisis de datos en Python?

Opciones:

a. Los outliers son valores atípicos que se encuentran en un conjunto de datos y se encuentran significativamente lejos de la mayoría de los otros puntos de datos. Estos valores inusuales pueden distorsionar las estadísticas y deben ser identificados y tratados adecuadamente.
b. Los outliers son una técnica de visualización en Python que se utiliza para mostrar gráficamente la relación entre dos variables numéricas en un conjunto de datos utilizando un gráfico de barras apiladas.
c. Los outliers son una técnica de regresión en Python que se utiliza para ajustar una línea recta a un conjunto de datos y predecir valores futuros basados en esa línea de regresión.
d. Los outliers son un método de agrupación en Python que se utiliza para dividir un conjunto de datos en grupos homogéneos con base en la similitud de características.

Pregunta 10

Supongamos que eres un analista de datos y tienes un conjunto de datos que contiene información sobre las ventas mensuales de una tienda durante un año. Tu objetivo es crear un gráfico que muestre la evolución de las ventas a lo largo del año. ¿Qué librería básica de graficación de datos en Python deberías utilizar para este propósito?

Opciones:

a. Pandas
b. Numpy
c. SciKit-Learn
d. Matplotlib

Pregunta 11

De las siguientes definiciones, ¿cuál consideras que define mejor lo que es el análisis exploratorio de datos?

Opciones:

a. Es una técnica que se utiliza exclusivamente en la etapa final del análisis de datos para generar informes visuales atractivos.
b. Es un método que se utiliza para eliminar datos atípicos y valores extremos de un conjunto de datos.
c. Es un enfoque que se utiliza para predecir resultados futuros sin examinar previamente los datos disponibles.
d. Es una fase inicial en el proceso de análisis de datos que se centra en la exploración y comprensión de los datos antes de aplicar modelos estadísticos.

Pregunta 12

Considerando el concepto de lo que es una variable de tipo nominal, elige la representación correcta:

Opciones:

a. Una variable que representa categorías o grupos sin un orden inherente, como colores de ojos.
b. Una variable que mide la magnitud o cantidad de un fenómeno, como la temperatura en grados Fahrenheit.
c. Una variable que registra datos en una escala numérica con un punto de inicio absoluto, como el tiempo en segundos.
d. Una variable que se utiliza para medir la relación de dependencia entre dos o más variables, como la correlación entre altura y peso.

Pregunta 13

¿Cuál es el propósito principal de utilizar tablas de frecuencias en el cálculo de resúmenes de variables categóricas?

Opciones:

a. Resaltar las tendencias en datos numéricos.
b. Resumir y mostrar la distribución de datos cualitativos.
c. Representar visualmente la distribución de datos cuantitativos.
d. Organizar información sobre variables continuas.

Pregunta 14

Como analista de datos necesitas utilizar técnicas multivariantes en la toma de decisiones en un presupuesto de mercadotecnia con la ayuda de la librería Matplotlib en Python. ¿Cuál es tu propósito principal?

Opciones:

a. Calcular el ROI (*Return on Investment*) de una única campaña de marketing.
b. Identificar relaciones y patrones complejos entre diferentes variables en los datos de *marketing*.
c. Crear gráficos de barras simples para representar los datos de ventas de productos.
d. Realizar análisis de texto en comentarios de clientes en línea.

Pregunta 15

¿Por qué es crucial realizar el preprocesamiento de datos antes de su análisis en la mayoría de los casos de estudio?

Opciones:

a. Porque los datos sin procesar contienen errores y ruido que pueden afectar la calidad de los resultados.
b. Para ahorrar tiempo y recursos, ya que el análisis directo de datos sin procesar es más rápido.
c. Para aumentar la complejidad de los datos y hacer que el análisis sea más interesante.
d. Porque el preprocesamiento no es relevante en el análisis de datos.

Pregunta 16

Elige el concepto que explique mejor lo que es la correlación lineal.

Opciones:

a. La correlación lineal se utiliza para medir la relación entre tres o más variables en lugar de solo dos.
b. La correlación lineal se refiere a la relación entre dos variables que son completamente independientes y no tienen ningún efecto entre sí.
c. La correlación lineal es una medida estadística que indica el grado de relación lineal entre dos variables. Cuando dos variables tienen una correlación lineal positiva, significa que a medida que una variable aumenta, la otra también lo hace de manera proporcional. Por ejemplo, si estudias más horas, tu calificación en un examen tiende a aumentar.
d. La correlación lineal se refiere a la relación entre variables categóricas en lugar de variables numéricas.

Pregunta 17

De los ejemplos siguientes, indica cual representa con claridad lo que es una correlación negativa desde el punto de vista estadístico.

Opciones:

a. Una correlación negativa es cuando dos variables están fuertemente relacionadas de manera directa, es decir, cuando una variable aumenta, la otra también aumenta. Ejemplo: A medida que la temperatura exterior aumenta, el consumo de helado aumenta.
b. Una correlación negativa es cuando dos variables no están relacionadas en absoluto, es decir, no hay ninguna relación entre ellas. Ejemplo: El color de los zapatos de una persona y su capacidad para resolver acertijos matemáticos.
c. Una correlación negativa es cuando dos variables están fuertemente relacionadas de manera inversa, es decir, cuando una variable aumenta, la otra disminuye. Ejemplo: A medida que la temperatura exterior disminuye, la demanda de café en una cafetería aumenta.
d. Una correlación negativa es cuando dos variables están relacionadas de manera aleatoria, es decir, su relación cambia constantemente sin un patrón discernible. Ejemplo: La cantidad de pájaros en un parque y el precio de las manzanas en un mercado.

Pregunta 18

¿Cuál de los siguientes valores del coeficiente de correlación de Pearson representa una correlación negativa y qué significa una correlación negativa?

Opciones:

a. 0.75
b. 0.20
c. -0.10
d. -0.95

Pregunta 19

¿Cuál de los siguientes métodos matemáticos es el más adecuado para medir la correlación entre dos variables en un estudio de investigación económica?

Opciones:

a. Coeficiente de correlación de Pearson (r).
b. Media aritmética.
c. Desviación estándar.
d. Coeficiente de correlación de Spearman.

Pregunta 20

En una tienda de ropa, se han registrado datos de ventas mensuales de dos productos: camisetas y pantalones, durante un período de un año. La gerencia está interesada en determinar si existe una relación entre la temperatura promedio mensual y, las ventas de camisetas y pantalones. Quieren saber si las ventas de estos productos aumentan o disminuyen en función de la temperatura.

Pregunta: ¿Qué método de análisis bivariado sería más adecuado para investigar la relación entre la temperatura promedio mensual y las ventas de camisetas y pantalones en esta tienda de ropa?

Opciones:

a. Regresión logística.
b. Análisis de correlación de Pearson.
c. Análisis de varianza (ANOVA).

d. Prueba t de Student.

Pregunta 21

¿Cuál es una de las razones más importantes para llevar a cabo un muestreo como base para un análisis estadístico?

Opciones:

a. Aumentar la precisión del análisis al recopilar datos de toda la población.
b. Facilitar la interpretación de los resultados al reducir la cantidad de datos recopilados.
c. Garantizar que los datos sean completamente aleatorios y sin sesgos.
d. Reducir la variabilidad en los datos al eliminar valores atípicos.

Pregunta 22

¿Cuál de las siguientes definiciones describe con precisión el concepto de población en el contexto de la estadística?

Opciones:

a. La población se limita a un grupo pequeño y específico de individuos que son seleccionados de manera aleatoria para representar a la totalidad de un conjunto mayor.
b. La población se refiere al conjunto completo de elementos o individuos que comparten una característica común y son el enfoque principal de un estudio estadístico.
c. La población se refiere a un grupo de individuos que participan en una encuesta o estudio de opinión y están dispuestos a responder preguntas sobre un tema específico.
d. La población se compone de todos los individuos de una región geográfica determinada, independientemente de cualquier característica compartida.

Pregunta 23

¿Cuál es una consecuencia del sesgo de selección en un estudio de investigación?

Opciones:

a. Aumenta la representatividad de la muestra.
b. Facilita la generalización de los hallazgos.
c. Disminuye la validez interna de los resultados.
d. Mejora la precisión de las mediciones.

Pregunta 24

De las características que definen mejor lo que es un muestreo aleatorio se puede decir que:

Opciones:

a. El muestreo aleatorio asegura que los elementos más comunes en la población tengan una mayor probabilidad de ser incluidos en la muestra.
b. El muestreo aleatorio implica seleccionar elementos de la población de manera predecible y sistemática.
c. En el muestreo aleatorio, los elementos de la población se eligen de forma intencional para representar las características específicas de interés.
d. El muestreo aleatorio garantiza que cada elemento de la población tenga la misma probabilidad de ser seleccionado en la muestra.

Pregunta 25

El objetivo principal de entrenar un modelo de análisis es:

Opciones:

a. Aprender a reconocer patrones y realizar predicciones precisas.
b. Identificar patrones en datos no relacionados.
c. Generar datos de entrenamiento aleatorios.
d. Aumentar el tamaño de la memoria del modelo.

Pregunta 26

Una de las principales razones del porqué es útil la visualización gráfica de datos puede decirse que es:

Opciones:

a. Muestra de forma atractiva los datos para presentarlos con máximo detalle.
b. La visualización gráfica de datos solo es relevante para empresas de tecnología o análisis de datos, no para otros sectores.
c. La visualización gráfica de datos es útil, pero se puede prescindir de ella si se cuenta con suficiente información numérica detallada.
d. Facilita la comprensión de tendencias, patrones y la identificación de valores atípicos.

Pregunta 27

¿Por qué es crucial que el formato de un gráfico contribuya a la claridad de la información en lugar de ser solo un elemento decorativo?

Opciones:

a. Los gráficos decorativos pueden ser útiles en presentaciones y documentos de *marketing*, donde la impresión visual es más importante que la información detallada.
b. El formato adecuado de un gráfico es esencial para la claridad de la información, ya que facilita la comprensión de los datos y la interpretación de tendencias.
c. La elección del formato de un gráfico depende principalmente de las preferencias personales del autor, ya que lo que es claro para uno puede no serlo para otro.
d. El formato decorativo de un gráfico puede ser preferible en ocasiones, ya que puede hacer que la información sea más atractiva visualmente y, por lo tanto,

más memorable.

Pregunta 28

¿Cuál de los siguientes no es un estilo de gráfico que ofrece la librería Seaborn en Python para la visualización de datos?

Opciones:

a. Gráfico de árbol.
b. Gráfico de dispersión.
c. Gráfico de violín.
d. Gráfico de barras.

Pregunta 29

Especifica la utilidad de los gráficos tipo treemap en el análisis de datos.

Opciones:

a. Para identificar patrones de distribución y proporción en conjuntos de datos jerárquicos.
b. Para representar datos en series temporales.
c. Para visualizar la relación entre dos variables numéricas.
d. Para mostrar la tendencia de datos categóricos en un período específico.

Pregunta 30

¿Cuál es la utilidad de los gráficos tipo scatterplot por categorías para el análisis de datos?

Opciones:

a. Para representar series de tiempo y analizar tendencias a lo largo del tiempo.
b. Para comparar la frecuencia de eventos en diferentes categorías de forma visual.
c. Para resaltar la correlación lineal entre dos variables numéricas en un solo gráfico.
d. Para resumir datos categóricos y mostrar la distribución de valores en diferentes grupos.

Pregunta 31

Explica en qué lugar requerirías realizar una prueba AB para evaluar la resistencia de materiales.

Opciones:

a. En un parque de diversiones.
b. En una biblioteca.
c. En una planta de energía nuclear.
d. En un laboratorio químico.

Pregunta 32

Define lo que es un “sujeto” en una prueba AB.

Opciones:

a. El sujeto en una prueba AB se refiere a una variable independiente utilizada para manipular las condiciones de la prueba y observar su efecto en el resultado.
b. El sujeto en una prueba AB se refiere a un conjunto de datos recopilados como resultado de la prueba, que luego se analizan para obtener conclusiones.
c. El sujeto en una prueba AB se refiere al participante o individuo que forma parte del estudio y es sometido a las diferentes condiciones de la prueba, permitiendo la comparación de resultados.
d. El sujeto en una prueba AB se refiere al investigador principal que diseña y realiza la prueba, supervisando todo el proceso.

Pregunta 33

¿Cuál de las siguientes afirmaciones describe con precisión lo que es el procesamiento del lenguaje natural (NLP)?

Opciones:

a. El procesamiento del lenguaje natural se refiere a la capacidad de las computadoras para entender y generar lenguaje humano de manera automatizada.
b. El procesamiento del lenguaje natural se limita a la traducción automática de un idioma a otro.
c. El procesamiento del lenguaje natural se centra únicamente en el análisis de palabras individuales y no tiene en cuenta la estructura gramatical.
d. El procesamiento del lenguaje natural se utiliza principalmente para el análisis de datos numéricos en textos.

Pregunta 34

¿Cuál es la mejor aproximación para comenzar el estudio del procesamiento del lenguaje natural?

Opciones:

a. Iniciar con un curso de programación para comprender la implementación de algoritmos NLP.
b. Comenzar con proyectos prácticos y el uso de bibliotecas NLP como NLTK o spaCy.
c. Investigar la historia y evolución de la inteligencia artificial antes de explorar NLP.
d. Aprender los fundamentos de la lingüística y gramática antes de abordar conceptos técnicos.

Pregunta 35

Por esta razón principalmente es que el procesamiento del lenguaje natural es uno de los retos más importantes de la inteligencia artificial:

Opciones:

a. El procesamiento del lenguaje natural es un reto importante para la inteligencia artificial porque requiere una gran cantidad de recursos computacionales.
b. El procesamiento del lenguaje natural es un reto importante para la inteligencia artificial debido a la falta de interés en aplicaciones prácticas.
c. El procesamiento del lenguaje natural es un reto importante para la inteligencia artificial debido a la ambigüedad y complejidad inherentes al lenguaje humano, lo que dificulta la comprensión automática de textos y conversaciones.
d. El procesamiento del lenguaje natural es un reto importante para la inteligencia artificial solo en entornos académicos.

Pregunta 36

Entre las siguientes afirmaciones, una describe con mayor precisión lo que es en concepto *Machine Learning.*

Opciones:

a. *Machine Learning.* es un enfoque de inteligencia artificial que solo se utiliza en aplicaciones de procesamiento de texto y lenguaje natural.
b. *Machine Learning.* se refiere a la programación manual de algoritmos complejos para que las computadoras realicen tareas específicas de manera eficiente.
c. *Machine Learning.* se basa en reglas de programación fijas que no pueden adaptarse ni mejorar con la experiencia.
d. *Machine Learning.* es un enfoque de inteligencia artificial que permite a las computadoras aprender y mejorar su rendimiento en tareas específicas a través de la experiencia y sin una programación explícita.

Pregunta 37

Elije un ejemplo de una actividad común que puedas realizar con ayuda de *Machine Learning.*

Opciones:

a. Traducción de idiomas.
b. Edición de imágenes.
c. Gestión de bases de datos.
d. Diseño gráfico.

Pregunta 38

¿Cuál de los siguientes algoritmos de aprendizaje supervisado es especialmente adecuado para la clasificación de datos en problemas con múltiples categorías?

Opciones:

a. Regresión Lineal.
b. SVM *(Support Vector Machine*).
c. Redes Neuronales.
d. Regresión Logística.

Pregunta 39

¿Cuál es el propósito principal de los algoritmos de aprendizaje no supervisado en el campo de la inteligencia artificial?

Opciones:

a. Clasificar automáticamente datos en categorías predefinidas.
b. Mejorar la precisión de los modelos de aprendizaje supervisado.
c. Encontrar patrones y estructuras ocultas en datos sin etiquetar.
d. Detectar errores en los conjuntos de datos etiquetados.

Pregunta 40

Imagina que estás trabajando en un proyecto de análisis de datos para una tienda en línea que vende una amplia variedad de productos. La tienda quiere segmentar a sus clientes en grupos para personalizar las ofertas y estrategias de marketing.

¿Qué enfoque de Machine Learning sería más adecuado para este caso?

Opciones:

a. Aprendizaje no supervisado.
b. Aprendizaje por refuerzo.
c. Aprendizaje semi-supervisado.
d. Aprendizaje supervisado.
