Clave de respuesta

FUNDAMENTOS DE MACHINE LEARNING_003V

Evaluación Formativa

¡Hola, equipo! 👋😄
Antes de que arranquen con su *misión evaluativa*, les dejo unas instrucciones importantes... pero sin tanto tecnicismo, al estilo *Fundamentos de Machine Learning* (porque todavía no somos robots... ¡pero vamos en camino! 🤖).

📌 **¿Qué está en juego?**
La evaluación sumativa vale el **10% de la cátedra.**

🧠 **Modo "Red neuronal activa"**
Esto es **individual**. Nada de redes colaborativas humanas en esta ocasión. Cada cerebro trabaja solito, como un buen modelo sin overfitting (sobre ajuste).

🚫 **Prohibido usar ayuditas externas**
No se permiten celulares, apuntes, ni herramientas de inteligencia artificial como ChatGPT, DeepSeek, Claude o cualquier otra IA que esté rondando por ahí. Hoy, ustedes son el algoritmo que tiene que hacer el *output* solito.

⏱️ **Tiempo total: 70 minutos**
Así que gestionen su tiempo como si estuvieran entrenando un modelo con límite de epochs (es como cuando un programa solo tiene cierto número de intentos para aprender algo). ¡No se pasen del tiempo!

↩️ **Importante: no se puede volver atrás**
Esto no es una interfaz interactiva con "Undo" (esa opción mágica de deshacer cuando uno se equivoca). Cada pregunta que avancen, quedó. ¡Lean bien antes de responder!

🗓️ **Y recuerden...**
La evaluación sumativa es el viernes 11 de abril de 2025. Marquen esa fecha como si fuera un hiperparámetro clave (esos valores importantes que definen cómo va a funcionar un modelo desde el inicio).

¡Éxito! 💥
Confíen en lo que han aprendido, y a demostrar que están listos para predecir, analizar y brillar 🌟.

Evaluación Formativa RA1 – Fundamentos de Machine Learning

Duración: 70 minutos
Formato: Individual, sin retroceso entre bloques
Objetivo: Comprender y reflexionar sobre los fundamentos de estadística descriptiva e inferencial, probabilidad y álgebra lineal. Conectar estos fundamentos con la importancia futura en Machine Learning.

BLOQUE 1: Tipos de datos y estadística descriptiva

Tipos de datos

Categóricos:

- Nominales (sin orden).
- Ordinales (con orden).

Numéricos:

- Discretos (enteros contables)
- Continuos (valores en un rango ininterrumpido)

Tablas de frecuencia

- Útiles para variables categóricas o numéricas discretas.
- Permiten contar cuántos registros caen en cada categoría o clase.
- En Machine Learning, sirven para conocer la **distribución** y la **frecuencia** de categorías antes de entrenar un modelo.

Medidas de tendencia central

- **Media**: promedio aritmético, sensible a valores extremos.
- **Mediana**: valor central; más robusta ante outliers.
- **Moda**: valor que se repite con mayor frecuencia (muy útil en datos categóricos).

Diferencia entre Estadística Descriptiva vs. Inferencial

- **Descriptiva**: se limita a describir o resumir los datos de una muestra.
- **Inferencial**: intenta generalizar conclusiones a toda una población, usando la muestra.

Aplicación en Machine Learning

- Exploración inicial de datos.
- Decisión sobre qué tipo de variable es cada atributo para elegir un método adecuado.

Pregunta 1

1 punto

Observa los siguientes ejemplos de “atributos” o columnas en un conjunto de datos:

1. Género (masculino, femenino, otro)
2. Temperatura (en grados Celsius)
3. Categoría del producto (básico, estándar, premium)
4. Cantidad de compras (número entero)

Asocia cada uno con el tipo de dato que le corresponde (categórico nominal, categórico ordinal, numérico discreto, numérico continuo) y explica **brevemente** por qué.

Ejemplo de una respuesta correcta

Ubicación:

- "1.1.1 Tipos de Datos y su aplicación.pptx" (diapositivas sobre “Tipos de datos”).

Respuesta breve:

- Género: Categórico nominal (no hay orden).
- Temperatura: Numérico continuo (puede tomar decimales).
- Categoría del producto: Categórico ordinal (básico < estándar < premium).
- Cantidad de compras: Numérico discreto (números enteros contables).

Con el siguiente enunciado responda la pregunta 2 y 3.

Imagina que tienes 500 registros de “Tipo de Mascota” (Perro, Gato, Otros) en una clínica veterinaria.

Pregunta 2

1 punto

¿Cómo usarías una **tabla de frecuencia** para mostrar la información?

Ejemplo de una respuesta correcta

Ubicación:

- "1.2.1 Estadísticas_Descriptiva_I.pptx" (sección “Tablas de Frecuencia”).

Respuesta breve:

- Se elabora una tabla donde cada fila representa un tipo de mascota (Perro, Gato, Otros) y, en la columna de frecuencia, se anota cuántos registros hay de cada tipo. Esto permite ver rápidamente cuántas observaciones hay por categoría.

Pregunta 3

1 punto

¿Por qué es útil contar la ocurrencia de cada categoría antes de aplicar un algoritmo de Machine Learning (por ejemplo, un clasificador)?

Ejemplo de una respuesta correcta

Ubicación:
- Igual que la pregunta 2 (“Tablas de Frecuencia” en "1.2.1 Estadísticas_Descriptiva_I.pptx").

Respuesta breve:
- Antes de entrenar un clasificador, conviene saber cuántos registros hay en cada clase (por ejemplo, “Perro”, “Gato”, “Otros”) para detectar desequilibrios y posibles problemas de sesgo en el modelo.

Pregunta 4

1 punto

¿Por qué la estadística descriptiva es la primera herramienta que usamos cuando iniciamos un proyecto de Machine Learning? Reflexiona sobre la **importancia de explorar y resumir** la información antes de cualquier modelo.

Ejemplo de una respuesta correcta

Ubicación:
- "1.2.1 Estadísticas_Descriptiva_I.pptx" (introducción a Estadística Descriptiva).

Respuesta breve:
- La estadística descriptiva ayuda a entender la forma, tendencia y posibles anomalías en los datos (valores extremos, distribución, etc.). Antes de entrenar un modelo de Machine Learning, es esencial conocer bien el conjunto de datos para tomar decisiones adecuadas (limpieza, transformación, etc.).

BLOQUE 2: Medidas de posición, dispersión y su visualización

Medidas de posición (tendencia central)

- Media, Mediana, Moda.

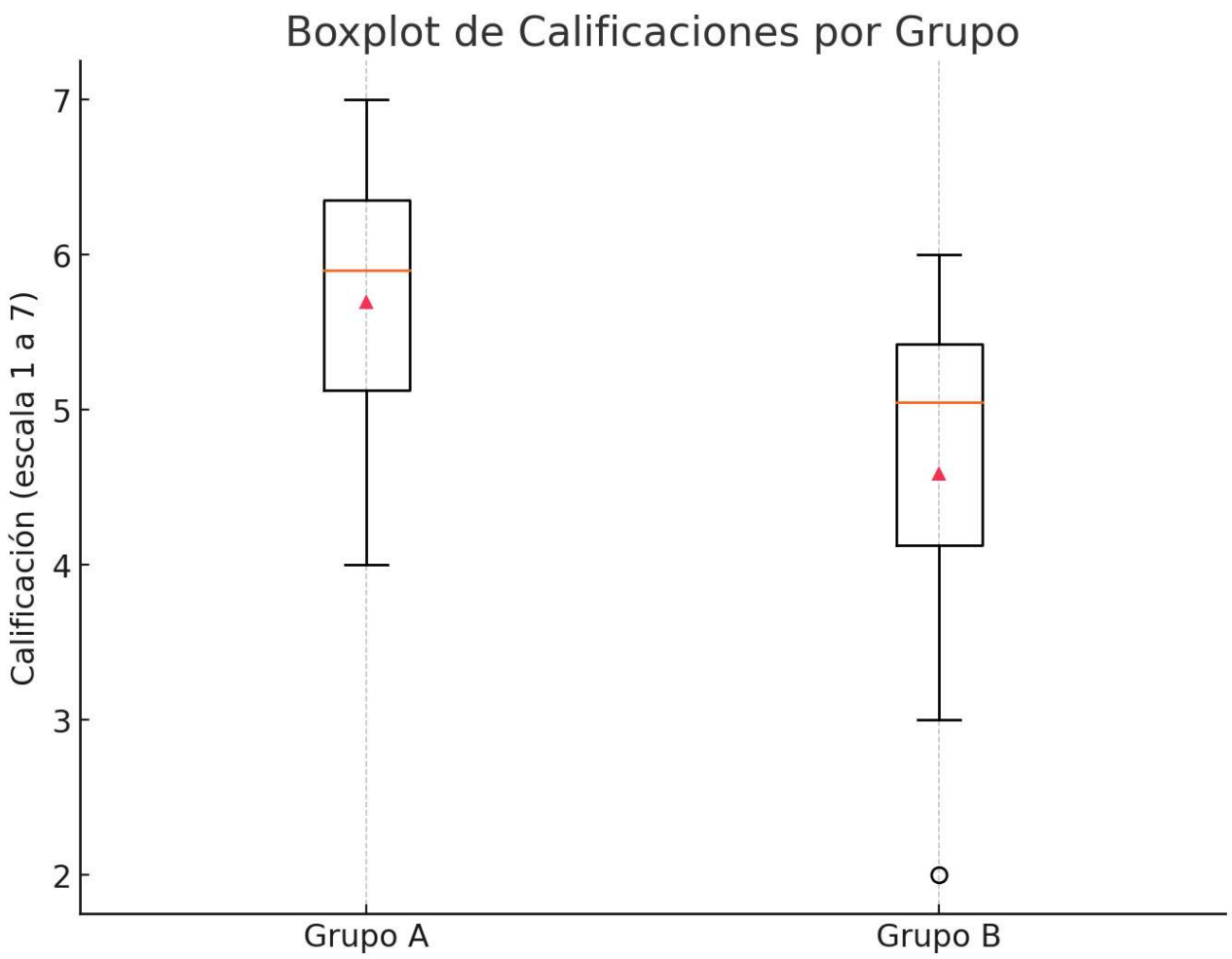
Medidas de dispersión

- Rango, desviación estándar, varianza, rango intercuartílico.
- Reflejan la variabilidad de los datos: mientras más grande, más se alejan del promedio.

Gráficos Boxplot

- Permiten comparar la **dispersión** y el **centro** entre grupos.

En un curso, se realizaron dos pruebas a diferentes grupos (A y B). A continuación, se muestra un boxplot (gráfico de cajas) con las calificaciones (rango 1 a 7).



Pregunta 5

2 puntos

Según este boxplot, ¿qué puedes decir sobre la **dispersión** de los puntajes en cada grupo? ¿Cuál grupo aparenta tener la **mediana** más alta? ¿Por qué es importante usar este tipo de visualizaciones antes de aplicar técnicas de Machine Learning?

Ejemplo de una respuesta correcta

Ubicación:

- "1.3.1 Medidas de dispersión.pptx" (contenido sobre dispersión) y complementariamente "1.2.1 Estadísticas_Descriptiva_I.pptx" (rango intercuartílico).

Respuesta breve:

- Un boxplot permite observar la mediana de cada grupo y qué tan dispersos están los datos. El grupo cuya caja (box) es más "ancha" o cuyos bigotes son más largos presenta mayor

dispersión. La mediana se identifica por la línea dentro de la caja. Es importante usar estas visualizaciones antes de aplicar ML para detectar outliers y entender la variabilidad.

Pregunta 6

2 puntos

Menciona **una ventaja** y **una desventaja** de la mediana respecto a la media cuando hay datos muy extremos. ¿Por qué a veces la mediana es más “representativa”? Conecta tu respuesta con un posible escenario donde en un dataset existan valores muy grandes por error de registro.

Ejemplo de una respuesta correcta

Ubicación:

- “1.2.1 Estadísticas_Descriptiva_I.pptx” (comparaciones entre media, mediana y moda).

Respuesta breve:

- Ventaja de la mediana: es más robusta a valores extremos; no se ve afectada por datos muy grandes o muy pequeños.
- Desventaja de la media: se desplaza mucho cuando hay outliers.
- La mediana puede ser más representativa en casos donde existan errores de registro con valores anormalmente altos o bajos.

BLOQUE 3: Probabilidad y eventos

Probabilidad en estadística

- Sirve para cuantificar la incertidumbre.
- Los eventos pueden ser **dependientes** o **independientes**.

Aplicación en Machine Learning

- Muchos algoritmos asumen independencia condicional de atributos.
- La forma en que modelamos la dependencia o independencia afecta la exactitud de predicciones.
- Distintas distribuciones (binomial, Poisson, normal) ayudan a describir situaciones, pero se requiere entender el concepto base de eventos y probabilidad.

Con el siguiente enunciado responde las preguntas 7, 8 y 9.

Piensa en **dos eventos** frecuentes en tu día a día (por ejemplo, “que llueva” y “que me quede dormido”), que podrían influirse entre sí o no.

Pregunta 7

1 punto

Describe brevemente esos dos eventos.

Ejemplo de una respuesta correcta

Ubicación:
- “1.4.1 Estadística_Descriptiva III.pptx” (sección de “Teoría de Probabilidades” y eventos independientes/dependientes).

Respuesta:
- Ejemplo de eventos cotidianos: “Que llueva” (A) y “Quedarme dormido” (B).

Pregunta 8

1 punto

Determina si son **independientes** o **dependientes** y por qué.

Ejemplo de una respuesta correcta

Ubicación:
- “1.4.1 Estadística_Descriptiva III.pptx” (sección de “Teoría de Probabilidades” y eventos independientes/dependientes).

Respuesta:
Si uno no influye en el otro (p. ej., que llueva no afecta mi alarma), son independientes. Si sí influyen (p. ej., si llueve fuerte, salgo más tarde y podría quedarme dormido), son dependientes.

Pregunta 9

1 punto

Explica, a nivel conceptual, por qué en Machine Learning es valioso entender esta distinción.

Ejemplo de una respuesta correcta

Ubicación:

- "1.4.1 Estadística_Descriptiva III.pptx" (sección de "Teoría de Probabilidades" y eventos independientes/dependientes).

Respuesta breve:
En Machine Learning, distinguir dependencia de independencia es crucial; muchos algoritmos (como Naive Bayes) hacen supuestos de independencia de atributos. Identificar relaciones reales entre variables mejora la exactitud de los modelos.

Pregunta 10

1 punto

Imagina un escenario de producción de galletas:
a. Cada galleta puede salir defectuosa (sí/no) con una probabilidad baja.
b. El número de clientes que llegan al día a la tienda puede variar bastante (hay un promedio, pero los datos fluctúan).
c. El **peso** de cada galleta está en torno a una media, con ligeras variaciones.

¿Qué tipo de distribución (binomial, Poisson o normal) **podría** describir cada una de estas variables? Justifica **sin usar fórmulas**, solamente la idea de por qué se ajusta mejor a cada situación.

Ejemplo de una respuesta correcta

Ubicación:
- "1.4.3 Distribución de Probabilidades.pptx" (distribuciones binomial, Poisson y normal).

Respuesta breve:
- a) Galleta defectuosa (sí/no) → se modela con distribución binomial (hay dos resultados posibles).
- b) Llegada de clientes al día, muy variable pero con un promedio → Poisson (sucesos que ocurren en un intervalo de tiempo).
- c) Peso de la galleta (fluye en torno a un promedio con variaciones suaves) → distribución normal.

BLOQUE 4: Comparando variables, enfoque descriptivo-inferencial

Estadística Inferencial

- Va más allá de describir la muestra; busca **inferir** conclusiones sobre la población.
- Ejemplo: ANOVA (Análisis de Varianza) para comparar medias entre varios grupos.

Correlación: Mide cuánto se mueven dos variables numéricas juntas (linealmente).

Regresión

- Para predecir o explicar una variable dependiente numérica desde otras variables (numéricas o categóricas).

Aplicación en Machine Learning

- ANOVA sirve para descubrir si un factor categórico (ej. tipo de tratamiento) influye en una variable continua (ej. resultados).
- La correlación o la regresión lineal son base de modelos predictivos (aunque en ML hay modelos más complejos, estos conceptos dan fundamentos).

Con el siguiente enunciado responde las preguntas 11 y 12.

En un estudio, se midió el rendimiento (variable numérica) de tres grupos de estudiantes que probaron **tres métodos de estudio distintos** (variable categórica).

Pregunta 11

1 punto

¿Por qué una técnica como **ANOVA** es más apropiada que una simple correlación en este caso?

Ejemplo de una respuesta correcta

Ubicación:
"1.3.3 Análisis de Varianza.pptx" (pasos de ANOVA).
Respuesta:

- Porque ANOVA sirve para comparar **tres grupos** con una **variable de números** (como notas). La correlación solo funciona si las dos cosas que se comparan **son números**, y aquí una es tipo de método (categoría). Por eso ANOVA es mejor en este caso.

Pregunta 12

1 punto

Si uno de los métodos sale mejor estadísticamente, ¿cómo podrías usar ese resultado en un proyecto futuro de Machine Learning (por ejemplo, recomendación de método)?

Ejemplo de una respuesta correcta

Ubicación:
"1.3.3 Análisis de Varianza.pptx" (pasos de ANOVA).
Respuesta (mejorada):
Si un método funciona mejor, entonces lo podemos usar en el futuro como el que **da mejores resultados**. También podemos guardar ese dato para que el computador **aprenda** cuál método ayuda más, y así pueda **predecir mejor** quién va a tener buen resultado.

Pregunta 13

1 punto

"Correlación" y "regresión" suelen confundirse, pero **no** son lo mismo. ¿Cómo le explicarías a un futuro compañero de equipo la diferencia conceptual? No uses fórmulas, solo la idea principal y **por qué** una no sustituye a la otra.

Ejemplo de una respuesta correcta

Ubicación:
- "1.5.1.Regresión y Correlación.pptx".

Respuesta:
- La correlación mide la relación lineal entre dos variables (qué tan unidas se mueven).
- La regresión se centra en predecir o explicar una variable (dependiente) a partir de otras (independientes) y produce una ecuación.
- No se sustituyen: correlación solo indica fuerza de asociación, la regresión define un modelo predictivo.

Pregunta 14

1 punto

Explica brevemente por qué **la estadística inferencial** es un paso adicional respecto a la descriptiva. Da **un ejemplo** en el que necesites hacer inferencia (no solo describir la muestra) y menciona por qué eso tiene relevancia cuando se entrena un modelo de Machine Learning.

Ejemplo de una respuesta correcta

Ubicación:
- "1.3.3 Análisis de Varianza.pptx" (introducción a estadística inferencial, pruebas de hipótesis).

Respuesta breve:
- La inferencia va más allá de describir la muestra y busca generalizar conclusiones a la población. Ej.: si solo describimos los datos de ciertos pacientes, hacemos estadística descriptiva; si queremos extrapolar a todos los pacientes del país y estimar un promedio o probar una hipótesis, hacemos estadística inferencial. En ML, usar inferencia ayuda a entender si los patrones hallados en la muestra se pueden generalizar.

BLOQUE 5: Álgebra lineal y representación de datos

Manejo de datos como matrices

- Filas = registros o muestras.
- Columnas = variables o atributos.
- Facilita operaciones vectorizadas (multiplicaciones de matrices) típicas en ML.

Sistemas de ecuaciones lineales

- Pueden no tener solución (restricciones inconsistentes), tener una única solución o infinitas soluciones.
- En ML, ajustar parámetros puede verse como “resolver” un sistema (ej. en regresión lineal simple).

Importancia en ML

- Muchos métodos se formulan con matrices (p. ej. redes neuronales, SVM, PCA).
- Entender sistemas lineales ayuda a interpretar por qué a veces no se puede entrenar un modelo (no hay solución viable) o hay infinitas soluciones (múltiples conjuntos de parámetros que encajan).

Pregunta 15

1 punto

¿Por qué crees que en Machine Learning es tan común representar la información en forma de “matrices” (filas como muestras, columnas como atributos)? ¿Qué ventajas aporta frente a manejar los datos de manera dispersa o desordenada?

Ejemplo de una respuesta correcta

Ubicación:

- “1.5.3 Álgebra Lineal.pptx” (matrices, sistemas de ecuaciones lineales).

Respuesta breve:
Porque es una forma **ordenada** de guardar los datos. Así, cada fila es una persona, cosa o ejemplo, y cada columna es una característica (como edad, nota, peso, etc.).
Esto hace que el computador pueda **trabajar más rápido y mejor**, haciendo cuentas con todos los datos al mismo tiempo, como si hiciera varias sumas o multiplicaciones de una sola vez.

Pregunta 16

1 punto

¿Qué significa que un sistema de ecuaciones lineales pueda **no tener solución**, tenga **una solución única** o **infinitas soluciones**? Relaciona esta idea con la posibilidad de que en ML a veces no podamos “ajustar” un modelo (datos contradictorios), otras veces lo ajustemos “justo”, y en otras hallemos múltiples configuraciones que funcionen igual de bien.

Ejemplo de una respuesta correcta

Ubicación:

- “1.5.3 Álgebra Lineal.pptx” (matrices, sistemas de ecuaciones lineales).

Respuesta breve:
Un sistema puede no tener solución si los datos **se contradicen**. Puede tener **una sola solución** si todo encaja justo. O puede tener **muchas soluciones** si hay varias formas que funcionan igual.
En Machine Learning pasa lo mismo:

- A veces los datos están mal y **no se puede entrenar el modelo**.
- Otras veces, hay **una forma exacta que funciona**.
- Y otras veces, **hay muchas formas diferentes** que dan buenos resultados.

Pregunta 17

2 puntos

Has visto cómo la estadística (descriptiva e inferencial), la probabilidad y el álgebra lineal se relacionan con la **comprensión** de datos. **¿De qué manera** crees que esto te facilitará el camino cuando más adelante estudies algoritmos de Machine Learning?

Ejemplo de una respuesta correcta

Respuesta:
Estas tres cosas ayudan mucho porque:

- La **estadística** sirve para **mirar y entender los datos**, ver si están bien o si hay cosas raras.
- La **probabilidad** ayuda a **tomar decisiones cuando hay duda**, como predecir algo que no se sabe con seguridad.
- El **álgebra lineal** se usa para que el computador **haga cálculos rápidos con muchos datos a la vez**.

Si se entienden bien estas ideas, seria más fácil cuando vea los algoritmos de Machine Learning, porque ya entenderé cómo funcionan por dentro.