EST. 2021 EMC
EDITORIAL MAR CARIBE

ESTADÍSTICA PARAMÉTRICA Y NO PARAMÉTRICA:

Aplicaciones en inteligencia artificial y machine learning

Escrito por:
María del Pilar Ríos García
Blanca Flor Robles Pastor
Maria Peregrina Cruzado Vallejos
Mariel del Rocío Chotón Calvo
Oscar Antonio Robles Villanueva
Julio Mariano Chávez Milla

www.editorialmarcaribe.es

ISBN: 978-9915-698-24-3

Estadística paramétrica y no paramétrica: Aplicaciones en inteligencia artificial y machine learning

María del Pilar Ríos García, Blanca Flor Robles Pastor, Maria Peregrina Cruzado Vallejos, Mariel del Rocío Chotón Calvo, Oscar Antonio Robles Villanueva, Julio Mariano Chávez Milla

© María del Pilar Ríos García, Blanca Flor Robles Pastor, Maria Peregrina Cruzado Vallejos, Mariel del Rocío Chotón Calvo, Oscar Antonio Robles Villanueva, Julio Mariano Chávez Milla, 2025

Primera edición: Agosto, 2025

Editado por:

Editorial Mar Caribe

www.editorialmarcaribe.es

Av. General Flores 547, Colonia, Colonia-Uruguay.

Diseño de portada: Yelitza Sánchez Cáceres

Libro electrónico disponible en:

https://editorialmarcaribe.es/ark:/10951/isbn.9789915698243

Formato: electrónico

ISBN: 978-9915-698-24-3

ARK: ark:/10951/isbn.9789915698243

Atribución/Reconocimiento-NoComercial 4.0 Internacional:

Los autores pueden autorizar al público en general a reutilizar sus obras únicamente con fines no lucrativos, los lectores pueden utilizar una obra para generar otra, siempre que se dé crédito a la investigación, y conceden al editor el derecho a publicar primero su ensayo bajo los términos de la licencia CC BY-NC 4.0.

Editorial Mar Caribe, firmante Nº 795 de 12.08.2024 de la Declaración de Berlín:

"... Nos sentimos obligados a abordar los retos de Internet como medio funcional emergente para la distribución del conocimiento. Obviamente, estos avances pueden modificar significativamente la naturaleza de la publicación científica, así como el actual sistema de garantía de calidad...." (Max Planck Society, ed. 2003., pp. 152-153).

Editorial Mar Caribe-Miembro de OASPA:

Como miembro de la Open Access Scholarly Publishing Association, apoyamos el acceso abierto de acuerdo con el código de conducta, transparencia y mejores prácticas de OASPA para la publicación de libros académicos y de investigación. Estamos comprometidos con los más altos estándares editoriales en ética y deontología, bajo la premisa de «Ciencia Abierta en América Latina y el Caribe».

Editorial Mar Caribe

Estadística paramétrica y no paramétrica: Aplicaciones en inteligencia artificial y machine learning

Colonia, Uruguay

2025

Sobre los autores y la publicación

María del Pilar Ríos García

https://orcid.org/0000-0002-0236-6810

Universidad Nacional de Tumbes, Perú

Blanca Flor Robles Pastor

https://orcid.org/0000-0002-7686-2825

Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas, Perú

Maria Peregrina Cruzado Vallejos

https://orcid.org/0000-0001-7809-4711

Universidad César Vallejo, Perú

Mariel del Rocío Chotón Calvo

https://orcid.org/0000-0001-6870-9268

Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas, Perú

Oscar Antonio Robles Villanueva

https://orcid.org/0000-0002-7540-5243

Universidad Nacional Mayor de San Marcos, Perú

Julio Mariano Chávez Milla

https://orcid.org/0000-0002-2335-8457

Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas, Perú

Resultado de la investigación del libro:

Publicación original e inédita, cuyo contenido es el resultado de un proceso de investigación realizado antes de su publicación, ha sido doble ciego de revisión externa por pares, el libro ha sido seleccionado por su calidad científica y porque contribuye significativamente al área del conocimiento e ilustra una investigación completamente desarrollada y completada. Además, la publicación ha pasado por un proceso editorial que garantiza su estandarización bibliográfica y usabilidad.

Sugerencia de citación: Ríos, M., Robles, B.F., Cruzado, M.P., Chotón, M., Robles, O.A., y Chávez, J.M. (2025). *Estadística paramétrica y no paramétrica: Aplicaciones en inteligencia artificial y machine learning*. Colonia del Sacramento: Editorial Mar Caribe. https://editorialmarcaribe.es/ark:/10951/isbn.9789915698243

Índice

Introducción 6

Capítulo I 8

Estadística Paramétrica y No Paramétrica: Claves para Potenciar la Inteligencia Artificial y el Machine Learning 8

1.1 Estadística paramétrica y no paramétrica: conceptos y aplicaciones 9

1.2 Estadística Paramétrica y No Paramétrica: Claves para su Aplicación en la Investigación Científica 16

1.3 Integración de Métodos de Inteligencia Artificial y Estadística Inferencial en la Investigación Científica: Un Enfoque Innovador 23

Capítulo II 32

Comparación de Medias y Análisis de Datos Binarios en Mediciones Dependientes: Fundamentos, Técnicas y Aplicaciones Prácticas 32

2.1 Fundamentos de la estadística en datos binarios 33

2.2 Prueba de Equivalencia (TOST) en Aprendizaje Automático: Fundamentos, Aplicaciones y Desafíos 39

2.3 Comparativa de la Prueba de Kruskal-Wallis y la Prueba de Mood en el Contexto del Aprendizaje Automático 45

Capítulo III 53

Comparación de Series de Datos Binarios: Varianzas y Proporciones en Análisis Estadístico 53

3.1 Proporciones en datos binarios 55

3.2 Comparativa de Pruebas Estadísticas en Machine Learning: McNemar y Q de Cochran 58

3.3 Comparativa de la Prueba de Fisher y la Prueba de Levene en Aprendizaje Automático: Aplicaciones, Limitaciones y Consideraciones Prácticas 65

Capítulo IV 72

Exploración de Pruebas de Asociación y Series de Tiempo: Fundamentos, Métodos y Aplicaciones 72

4.1 Pruebas de asociación..... 73

4.2 Análisis de Homogeneidad en Series Temporales y Aplicación del Modelo Multiplicativo Estacional de Holt-Winters 78

4.3 metodología para Ajustar Modelos ARIMA en Series Temporales.. 84

Conclusión..... 91

Bibliografía..... 93

Introducción

La estadística paramétrica es un enfoque que se basa en la suposición de que los datos provienen de una distribución específica, generalmente una distribución normal. Este enfoque se caracteriza por la utilización de parámetros, como la media y la varianza, para describir y modelar los datos. Sin embargo, es fundamental considerar sus limitaciones y evaluar si son adecuadas para el problema específico que se está abordando.

En tanto, la estadística no paramétrica se refiere a un conjunto de métodos estadísticos que no asumen una distribución específica para los datos, esto la convierte en una herramienta especialmente valiosa en contextos donde las suposiciones de la estadística paramétrica no se cumplen, lo que es común en muchos escenarios del mundo real, incluidos aquellos en inteligencia artificial (IA) y machine learning (ML).

En este libro se sistematizan los fundamentos, técnicas y aplicaciones de ambas corrientes dada la creciente disponibilidad de grandes volúmenes de datos y la necesidad de análisis más sofisticados para impulsar la investigación en nuevas técnicas que combinen enfoques paramétricos y no paramétricos; esto podría resultar en modelos más robustos y versátiles, capaces de adaptarse a la variabilidad inherente de los datos del mundo real. Por lo que la inteligencia artificial probablemente se beneficiará de un enfoque híbrido, donde las fortalezas de ambos tipos de estadística se aprovechen para mejorar la precisión y la capacidad predictiva de los modelos.

Por lo tanto, la elección entre estadística paramétrica y no paramétrica no es una cuestión de preferencia, depende en gran medida de la naturaleza de los datos y del problema en cuestión. Si bien los métodos paramétricos

pueden ofrecer soluciones rápidas y efectivas cuando se cumplen sus supuestos, los métodos no paramétricos brindan flexibilidad y robustez en contextos más variados. En última instancia, el éxito en la aplicación de la estadística en IA y ML radica en comprender las características de estos datos y seleccionar el enfoque adecuado que maximice la efectividad del análisis.

La convergencia de la inteligencia artificial y la estadística inferencial ha abierto nuevas vías para la investigación y la innovación, la capacidad de la IA para analizar datos representativos o no y reconocer patrones complejos se complementa con las herramientas de la estadística inferencial, que permiten validar hipótesis y estimar parámetros poblacionales. Esta sinergia mejora la precisión de los resultados, y acelera el proceso de descubrimiento científico. La inteligencia artificial proporciona técnicas avanzadas que pueden manejar y extraer información útil de conjuntos de datos masivos, mientras que la estadística inferencial ofrece un marco riguroso para interpretar esos resultados y establecer conclusiones válidas.

Ahora bien, hay una realidad y es en el ámbito del aprendizaje automático, pues, la correcta aplicación de pruebas estadísticas es fundamental para garantizar la validez y la fiabilidad de los modelos construidos. En este sentido, el objetivo de investigación es proporcionar una visión integral sobre las diferencias y aplicaciones de la estadística paramétrica y no paramétrica en el contexto de la inteligencia artificial y el machine learning. A través de un análisis detallado de sus características, ventajas y desventajas, se espera que los lectores comprendan cuándo y cómo utilizar cada enfoque estadístico en sus proyectos. Además, se abordarán las implicaciones futuras del uso de estas técnicas en el campo de la IA, lo que permitirá a los profesionales estar mejor equipados para enfrentar los desafíos que presentan los datos en constante evolución.

Capítulo I

Estadística Paramétrica y No Paramétrica: Claves para Potenciar la Inteligencia Artificial y el Machine Learning

La estadística juega un papel fundamental en el desarrollo y la implementación de algoritmos de inteligencia artificial (IA) y machine learning (ML), es decir, proporciona las herramientas necesarias para modelar la incertidumbre y tomar decisiones informadas basadas en datos. La estadística se divide en dos ramas principales: la estadística paramétrica y la no paramétrica, cada una con características y aplicaciones específicas. La estadística paramétrica se basa en suposiciones sobre la distribución de los datos, como la normalidad, y utiliza estas suposiciones para estimar parámetros de la población. En contraste, la estadística no paramétrica no requiere estas suposiciones y se centra en las propiedades intrínsecas de los datos, lo que la hace más flexible en ciertas situaciones.

En el contexto de la IA y el ML, la estadística es esencial para construir modelos predictivos y de clasificación que sean precisos y confiables. Los algoritmos de machine learning aprenden patrones a partir de datos históricos, y la estadística proporciona el marco necesario para entender y validar esos patrones. Los análisis estadísticos permiten a académicos y desarrolladores evaluar la eficacia de sus modelos, detectar sobreajustes y optimizar el rendimiento general de los sistemas de IA.

1.1 Estadística paramétrica y no paramétrica: conceptos y aplicaciones

Los métodos estadísticos paramétricos se fundamentan en varios supuestos clave, que incluyen:

- Normalidad de los datos: Implica que los datos están distribuidos de manera simétrica alrededor de la media, lo cual es crucial para la validez de muchas pruebas estadísticas.

- Homogeneidad de la varianza: Sugiere que la variabilidad dentro de cada grupo es similar.

- Independencia de las observaciones: Supone que las observaciones no están correlacionadas entre sí.

Cuando estos supuestos se cumplen, los métodos paramétricos tienden a ser más poderosos y eficientes, permitiendo estimaciones más precisas y, en general, mejores resultados en la inferencia estadística. Uno de los campos más destacados donde se aplica la estadística paramétrica es en los modelos de regresión. La regresión lineal, así como, es un método paramétrico que modela la relación entre una variable dependiente y una o más variables independientes. Este enfoque es fundamental en la predicción y análisis de datos en IA y ML, ya que permite identificar y cuantificar relaciones estadísticas, así como realizar inferencias sobre nuevas observaciones.

En el contexto del machine learning, la regresión lineal se utiliza a menudo como un modelo base debido a su simplicidad y facilidad de interpretación. A pesar de que existen técnicas más complejas, la regresión lineal proporciona una referencia útil para evaluar el rendimiento de modelos más sofisticados.

Otro ámbito de aplicación de la estadística paramétrica es en las redes neuronales. Aunque estos modelos son inherentemente no lineales y complejos, la parametrización de sus pesos y sesgos se basa en principios estadísticos. Durante el proceso de entrenamiento, las redes neuronales ajustan estos parámetros para minimizar la función de pérdida, que cuantifica la diferencia entre las predicciones del modelo y los valores reales.

Las técnicas paramétricas son fundamentales en la regularización de redes neuronales, donde se aplican métodos como Lasso o Ridge para evitar el sobreajuste. Estos enfoques incorporan penalizaciones al ajuste de los parámetros del modelo, favoreciendo soluciones más generales que se desempeñan mejor en datos no vistos (Molina, 2025). La estadística paramétrica juega un papel crucial en el desarrollo y la implementación de modelos en inteligencia artificial y machine learning. Los supuestos que la sustentan permiten obtener estimaciones más precisas y eficientes, especialmente en aplicaciones como la regresión y las redes neuronales, donde la parametrización es esencial para el aprendizaje y la predicción.

La estadística no paramétrica se refiere a un conjunto de métodos estadísticos que no hacen supuestos estrictos sobre la forma de la distribución de los datos. A diferencia de la estadística paramétrica, que se basa en parámetros específicos (como la media y la desviación estándar), la estadística no paramétrica se centra en la clasificación y el orden de los datos, lo que la hace especialmente útil en situaciones donde los datos no cumplen con los supuestos necesarios para aplicar técnicas paramétricas.

Para Flores et al (2017), la estadística no paramétrica permite realizar análisis sin requerir que los datos sigan una distribución normal o que tengan varianzas homogéneas, esto la convierte en una herramienta valiosa cuando

se trabaja con datos ordinales o cuando las muestras son pequeñas y pueden no reflejar adecuadamente la población. Entre sus características más destacadas se incluyen la flexibilidad y la capacidad de manejar datos con características no lineales o no distribuidas de manera estándar.

Algunos de los métodos más comunes de estadística no paramétrica incluyen la prueba de Wilcoxon, la prueba de Kruskal-Wallis y la regresión de rangos. Estos métodos se basan en la comparación de las posiciones relativas de los datos en lugar de en los valores absolutos, lo que permite obtener conclusiones válidas incluso cuando los datos son difíciles de modelar.

La estadística no paramétrica es especialmente efectiva en el ámbito de la clasificación de datos, donde se busca asignar etiquetas a instancias basándose en sus características. En concreto, en problemas de clasificación binaria, métodos como el clasificador de k-vecinos más cercanos (k-NN) utilizan distancias entre puntos en un espacio multidimensional para determinar a qué clase pertenece un nuevo punto. Este enfoque no requiere suposiciones sobre la distribución de los datos, lo que le permite adaptarse a diferentes tipos de distribuciones.

Además, técnicas como los árboles de decisión y los métodos de agrupamiento (clustering) también se benefician de enfoques no paramétricos. Estos métodos permiten segmentar datos en grupos significativos sin necesidad de realizar suposiciones sobre la estructura subyacente de los datos, lo que los hace ideales para aplicaciones en áreas como la segmentación de clientes y el análisis de mercado. En el contexto del aprendizaje no supervisado, la estadística no paramétrica ofrece herramientas poderosas para el análisis de datos sin etiquetas. Algoritmos como el análisis de componentes principales (PCA) y el t-SNE (t-distributed Stochastic

Neighbor Embedding) permiten visualizar y reducir la dimensionalidad de los datos de manera efectiva, facilitando la identificación de patrones y relaciones intrínsecas.

La ventaja de estos métodos es que no requieren que los datos se ajusten a una distribución particular, lo que los hace robustos frente a las variaciones y peculiaridades que pueden presentarse en conjuntos de datos reales. Además, las técnicas de agrupamiento jerárquico, que se basan en la similitud entre las observaciones, pueden adaptarse fácilmente a diferentes estructuras de datos, lo que permite descubrir grupos interesantes sin la necesidad de predefinir el número de clústeres.

La estadística no paramétrica juega un papel crucial en la inteligencia artificial y el machine learning, ofreciendo métodos flexibles y robustos que son especialmente útiles en contextos donde los supuestos paramétricos no son válidos. Su capacidad para clasificar y analizar datos sin restricciones de distribución las convierte en herramientas indispensables en la era de los datos. La elección entre métodos estadísticos paramétricos y no paramétricos es una decisión crucial en el ámbito de la inteligencia artificial (IA) y el aprendizaje automático (ML), dado que cada enfoque presenta características particulares que pueden influir en el rendimiento de los modelos.

Los métodos paramétricos, como su nombre indica, dependen de un número limitado de parámetros que describen la población de datos. Esto simplifica el análisis, ya que estos métodos pueden ser más rápidos y requieren menos datos para estimar los parámetros. A saber, en modelos de regresión lineal, se asume que la relación entre las variables es lineal y que los errores siguen una distribución normal. Esta simplificación permite realizar inferencias más fácilmente y es recomendable cuando se cumplen los

supuestos subyacentes. Sin embargo, las desventajas de los métodos paramétricos son evidentes cuando los supuestos no se cumplen. Si la relación entre las variables no es lineal o si los datos presentan distribuciones no normales, los modelos paramétricos pueden producir resultados engañosos o inexactos. Esto puede llevar a una falta de generalización en situaciones del mundo real.

Por otro lado, los métodos no paramétricos son más flexibles, ya que no asumen ninguna forma específica de la distribución de los datos. Esto permite que se adapten mejor a una variedad de situaciones, especialmente cuando se trabaja con datos no estructurados o en contextos donde los supuestos paramétricos no son válidos. Sin embargo, esta flexibilidad puede venir a costa de una mayor complejidad computacional y la necesidad de un mayor volumen de datos para obtener resultados fiables. Además, los métodos no paramétricos pueden ser menos eficientes en términos de interpretación y pueden requerir técnicas más avanzadas para su implementación.

La elección entre métodos paramétricos y no paramétricos depende en gran medida del contexto del problema y la naturaleza de los datos. Se puede citar, en escenarios donde se dispone de grandes conjuntos de datos y se pueden hacer supuestos razonables sobre la distribución de los datos, como en el análisis de datos de ventas o de comportamiento del cliente, los métodos paramétricos como la regresión lineal o la regresión logística son opciones adecuadas y eficientes.

En contraste, en situaciones donde los datos son escasos, ruidosos o no siguen una distribución específica, como en la clasificación de imágenes o en el procesamiento del lenguaje natural, los métodos no paramétricos como los

árboles de decisión o los algoritmos basados en vecinos más cercanos (K-NN) pueden ofrecer mejores resultados. Estos métodos permiten capturar patrones complejos sin imponer restricciones previas sobre la forma de los datos. La precisión de los modelos en IA y ML puede variar significativamente según el enfoque estadístico utilizado. Los métodos paramétricos tienden a ser más precisos cuando los supuestos son válidos, lo que se traduce en una mejor capacidad predictiva. Sin embargo, en situaciones donde los supuestos son violados, la precisión puede disminuir drásticamente (Alexander et al., 2015).

Por otro lado, los métodos no paramétricos, al ser más adaptativos, pueden lograr una mayor precisión en situaciones complejas y no lineales, aunque a menudo requieren más tiempo de entrenamiento y una mayor cantidad de muestras para evitar el sobreajuste. La clave está en balancear la complejidad del modelo con la cantidad y calidad de los datos disponibles, así como en realizar pruebas exhaustivas para determinar qué enfoque proporciona mejores resultados en un caso específico. Tanto los métodos paramétricos como los no paramétricos tienen su lugar en el ámbito de la estadística aplicada a la inteligencia artificial y el aprendizaje automático. La elección del método adecuado dependerá de las características de los datos y del problema específico a resolver, lo que resalta la importancia de una comprensión profunda de los fundamentos estadísticos en estos campos.

La estadística paramétrica, con sus supuestos específicos sobre la distribución de los datos, ha demostrado ser fundamental en la construcción de modelos de regresión y en el diseño de redes neuronales, permitiendo la formulación de inferencias precisas y la optimización de parámetros. Por otro lado, la estadística no paramétrica, al no depender de supuestos estrictos sobre la distribución de los datos, ha encontrado su lugar en aplicaciones de

clasificación y en algoritmos de aprendizaje no supervisado, donde la flexibilidad y la adaptabilidad son esenciales.

La elección entre métodos paramétricos y no paramétricos no es trivial y debe basarse en la naturaleza de los datos, el contexto del problema y los objetivos específicos del análisis. Entre tanto que los métodos paramétricos pueden ofrecer soluciones más eficientes cuando se cumplen sus supuestos, los enfoques no paramétricos son herramientas valiosas en situaciones donde la flexibilidad y la robustez son necesarias. Esta decisión impacta directamente en la precisión y efectividad de los modelos desarrollados, lo que subraya la importancia de una comprensión clara de las características de cada enfoque.

Se prevé que la integración de métodos estadísticos avanzados con técnicas emergentes, como el aprendizaje profundo y la inteligencia artificial explicativa, continúe transformando el panorama. Además, el aumento de la disponibilidad de grandes volúmenes de datos impulsará la necesidad de enfoques estadísticos más sofisticados que puedan manejar la complejidad y la heterogeneidad de la información.

Asimismo, la creciente interdisciplinariedad en la investigación y aplicación de la IA y el ML sugiere que la colaboración entre estadísticos, científicos de datos y expertos en dominio específico será crucial para el desarrollo de métodos más robustos y adaptativos. Por ende, la educación y la formación en estadística aplicada a estas áreas serán fundamentales para preparar a la próxima generación de profesionales que enfrentarán los desafíos y oportunidades que la inteligencia artificial tiene por delante. La estadística, en sus formas paramétricas y no paramétricas, seguirá desempeñando un papel esencial en el desarrollo y la mejora de los modelos

de IA y ML, facilitando la toma de decisiones informadas y el avance en diversas aplicaciones que impactan nuestras vidas cotidianas.

1.2 Estadística Paramétrica y No Paramétrica: Claves para su Aplicación en la Investigación Científica

La estadística es una herramienta fundamental en la investigación científica, ya que permite a los académicos recopilar, analizar e interpretar datos de manera efectiva. Dentro del ámbito estadístico, existen dos enfoques principales: la estadística paramétrica y la no paramétrica. La estadística paramétrica se basa en ciertos supuestos sobre la distribución de los datos, como la normalidad y la homogeneidad de varianzas, este enfoque permite aplicar diversas técnicas analíticas, como el análisis de varianza (ANOVA) y la regresión lineal, que son altamente potentes y eficaces cuando se cumplen sus supuestos.

Por otro lado, la estadística no paramétrica no asume ninguna distribución específica de los datos, lo que la convierte en una opción más flexible y robusta en situaciones donde no se pueden cumplir los supuestos paramétricos; entre sus técnicas se encuentran las pruebas de rangos, como la prueba de Mann-Whitney y la prueba de Kruskal-Wallis, que son especialmente útiles en estudios con muestras pequeñas o datos ordinales (Roig, 2019).

La estadística desempeña un papel crucial en cada etapa de la investigación científica. Desde el diseño del estudio y la selección de muestras adecuadas, hasta el análisis de los resultados y la formulación de conclusiones, la estadística proporciona un marco para tomar decisiones informadas. Sin un enfoque estadístico riguroso, los académicos corren el riesgo de caer en sesgos o malinterpretaciones de los datos.

Además, la capacidad de generalizar resultados de una muestra a una población más amplia, evaluar la significancia de los hallazgos y establecer relaciones entre variables son aspectos que dependen en gran medida de métodos estadísticos bien fundamentados. La estadística paramétrica se basa en ciertos supuestos sobre la distribución de los datos, lo que le permite realizar inferencias más precisas y potentes en comparación con los métodos no paramétricos.

El análisis de varianza, comúnmente conocido como ANOVA, es una técnica estadística utilizada para comparar las medias de tres o más grupos. Este método permite determinar si hay diferencias significativas entre las medias de los grupos bajo estudio. ANOVA se basa en el supuesto de que los datos siguen una distribución normal y que las varianzas de los grupos son homogéneas. Un uso típico de ANOVA se encuentra en experimentos donde se evalúa el efecto de diferentes tratamientos sobre una variable dependiente, como en estudios farmacológicos, donde se comparan los efectos de varios medicamentos sobre la presión arterial.

La regresión lineal es otra herramienta fundamental de la estadística paramétrica que se utiliza para modelar la relación entre una variable dependiente y una o más variables independientes. Este método permite prever valores de la variable dependiente y entender cómo cambian con respecto a las variables independientes. A saber, en investigaciones sobre el impacto de la educación en los ingresos, se puede utilizar la regresión lineal para cuantificar cómo un aumento en los años de educación se asocia con un incremento en los ingresos anuales. La regresión lineal también permite

realizar inferencias sobre la relación entre variables, lo que puede ser valioso para la formulación de políticas.

Las pruebas t son procedimientos estadísticos que se utilizan para determinar si hay diferencias significativas entre las medias de dos grupos. Estas pruebas son particularmente útiles en estudios donde se desea comparar dos condiciones o grupos, como un grupo de control y un grupo experimental. Existen diferentes tipos de pruebas t, como la prueba t de Student para muestras independientes y la prueba t de muestras pareadas. Así como, en un estudio que evalúa el efecto de un programa de intervención sobre el nivel de ansiedad, se podría utilizar una prueba t para comparar los niveles de ansiedad antes y después de la intervención entre los participantes.

La estadística paramétrica ofrece herramientas poderosas que permiten a los académicos analizar y comprender datos complejos, siempre y cuando se cumplan los supuestos necesarios para su correcta aplicación. Las técnicas mencionadas, como ANOVA, la regresión lineal y las pruebas t, son fundamentales en el ámbito científico y contribuyen significativamente a la validez y robustez de los hallazgos de investigación. La estadística no paramétrica se presenta como una herramienta invaluable en el análisis de datos, especialmente cuando las suposiciones de normalidad y homogeneidad de varianzas que sustentan los métodos paramétricos no se cumplen. Este tipo de estadística es particularmente útil en situaciones donde los datos son ordinales, o cuando se dispone de muestras pequeñas.

La prueba de Mann-Whitney, también conocida como prueba U de Mann-Whitney, es un método utilizado para comparar dos grupos independientes cuando los datos no siguen una distribución normal. Esta prueba evalúa si hay diferencias significativas en las medianas de ambos

grupos, lo que la convierte en una opción ideal en estudios de ciencias sociales, psicología y medicina (Hard, 2001). Pongamos el caso de, se puede utilizar para analizar la eficacia de dos tratamientos diferentes en pacientes, donde los resultados son medidos en una escala ordinal.

La prueba de Wilcoxon, o prueba de rangos con signo de Wilcoxon, es empleada para comparar dos muestras relacionadas o emparejadas. Esta prueba es útil en estudios donde se desea evaluar los efectos de un tratamiento antes y después de su aplicación. En concreto, en un estudio clínico que examine el impacto de una intervención sobre la presión arterial, los valores de presión de los pacientes pueden ser medidos antes y después del tratamiento. La prueba de Wilcoxon permite determinar si hay una diferencia significativa en las mediciones emparejadas, sin asumir que los datos siguen una distribución normal.

La prueba de Kruskal-Wallis es una extensión de la prueba de Mann-Whitney que permite comparar más de dos grupos independientes. Esta prueba es útil en investigaciones en las que se analizan múltiples tratamientos o condiciones y se busca determinar si al menos uno de los grupos difiere significativamente de los demás. En particular, en un ensayo clínico que evalúe diferentes dosis de un medicamento, la prueba de Kruskal-Wallis puede ayudar a identificar si alguna de las dosis produce efectos diferentes en comparación con las otras. Al ser una prueba no paramétrica, es ideal para datos que no cumplen con los supuestos necesarios de las pruebas paramétricas.

Las aplicaciones de la estadística no paramétrica son diversas y cruciales en la investigación científica, especialmente en contextos donde los datos no se ajustan a las condiciones requeridas para los métodos

paramétricos. Estas pruebas permiten a los académicos obtener conclusiones significativas y válidas, asegurando la robustez y la fiabilidad de sus hallazgos. La elección entre estadística paramétrica y no paramétrica es un aspecto crucial en el diseño y análisis de estudios científicos. Ambos enfoques tienen sus propias características, ventajas y desventajas, que los hacen adecuados para diferentes tipos de datos y situaciones de investigación.

La estadística paramétrica se basa en supuestos sobre la distribución de los datos, generalmente asumiendo que siguen una distribución normal. Esto permite realizar análisis más potentes y precisos, especialmente cuando se cumplen estos supuestos. Entre las ventajas de la estadística paramétrica se encuentran:

i. *Mayor potencia estadística:* Cuando los supuestos son válidos, las pruebas paramétricas pueden detectar diferencias con mayor eficacia.
ii. *Facilidad de interpretación:* Muchos de los resultados estadísticos son más intuitivos y están bien establecidos en la literatura.
iii. *Modelado más flexible:* Permite el uso de modelos más complejos, como la regresión lineal múltiple, que pueden incluir múltiples variables independientes.

Sin embargo, la estadística paramétrica también tiene desventajas:

i. *Dependencia de supuestos:* Si los datos no cumplen con los supuestos de normalidad o homogeneidad de varianzas, los resultados pueden ser engañosos.
ii. *Limitaciones en tipos de datos:* No es adecuada para datos ordinales o nominales, donde la información sobre el orden o la categoría es importante.

Por otro lado, la estadística no paramétrica no se basa en supuestos estrictos sobre la distribución de los datos, lo que la hace más versátil en ciertos contextos. Sus ventajas incluyen:

i. *Flexibilidad:* Puede aplicarse a datos que no cumplen con los supuestos de normalidad, como datos ordinales o distribuciones sesgadas.
ii. *Robustez:* Es menos sensible a valores atípicos y distribuciones irregulares, lo que puede hacer que los resultados sean más fiables en algunos casos.
iii. *Uso en muestras pequeñas:* Las pruebas no paramétricas son a menudo más apropiadas para muestras pequeñas, donde los supuestos de las pruebas paramétricas pueden no ser válidos.

Sin embargo, la estadística no paramétrica también presenta desventajas:

i. *Menor potencia:* En general, las pruebas no paramétricas pueden ser menos potentes que sus equivalentes paramétricos cuando se cumplen los supuestos.
ii. *Interpretación más compleja:* Algunos resultados pueden ser más difíciles de interpretar, ya que no se basan en medidas como la media o la varianza.

La elección entre estadística paramétrica y no paramétrica depende, en gran medida, de la naturaleza de los datos y los objetivos del análisis. Si los datos son continuos y cumplen con las condiciones de normalidad y homogeneidad de varianzas, las pruebas paramétricas son generalmente preferibles. Por el contrario, si los datos son ordinales, categóricos, o si hay dudas sobre la normalidad, las pruebas no paramétricas son más adecuadas.

Además, es importante considerar el tamaño de la muestra. En muestras pequeñas, la estadística no paramétrica puede ser la mejor opción para evitar conclusiones erróneas. En contraste, con muestras grandes, donde el teorema central del límite sugiere que la distribución de la media se aproxima a la normalidad, las pruebas paramétricas pueden ser más ventajosas.

Existen muchos estudios en diferentes campos de la investigación que emplean tanto estadística paramétrica como no paramétrica. Para ilustrar, en la investigación médica, un estudio podría usar ANOVA para comparar la eficacia de varios tratamientos en grupos grandes, y, al mismo tiempo, utilizar la prueba de Mann-Whitney para analizar datos de un grupo pequeño de pacientes donde los supuestos paramétricos no son válidos. En la psicología, un investigador podría emplear regresión lineal para examinar la relación entre la autoestima y el rendimiento académico, mientras que, en otro análisis, podría optar por la prueba de Wilcoxon para evaluar diferencias en puntuaciones de ansiedad entre grupos de tratamiento.

La elección entre estadística paramétrica y no paramétrica no es simplemente una cuestión de preferencia; depende del contexto, los datos y los objetivos específicos del estudio. La comprensión de las ventajas y desventajas de cada enfoque permite a los académicos tomar decisiones informadas que mejoren la validez y la fiabilidad de sus resultados.

La estadística paramétrica, con métodos como el ANOVA y la regresión lineal, permite realizar inferencias basadas en supuestos sobre la distribución de los datos, ofreciendo herramientas poderosas para comparar grupos y estudiar relaciones entre variables; por otro lado, la estadística no paramétrica, a través de pruebas como Mann-Whitney y Kruskal-Wallis,

proporciona alternativas robustas cuando los datos no cumplen con los supuestos necesarios para los métodos paramétricos (Burbano et al., 2022).

La elección del método estadístico adecuado es crucial en cualquier investigación, ya que puede influir significativamente en los resultados y sus interpretaciones. La comprensión de las ventajas y desventajas de cada enfoque permite a los académicos seleccionar la técnica más apropiada, garantizando así la validez de sus conclusiones. La integración de ambas metodologías en un mismo estudio puede ofrecer una visión más completa y enriquecer el análisis de los datos.

Mirando hacia el futuro, es esencial que los académicos continúen desarrollando y adaptando métodos estadísticos para abordar las complejidades del mundo real. La evolución de las técnicas estadísticas, junto con el crecimiento de la ciencia de datos y el aprendizaje automático, promete abrir nuevas vías para la investigación y el análisis. La estadística seguirá siendo un pilar fundamental en la ciencia, y su correcta aplicación será clave para avanzar en el conocimiento y la comprensión de fenómenos complejos.

1.3 Integración de Métodos de Inteligencia Artificial y Estadística Inferencial en la Investigación Científica: Un Enfoque Innovador

En la última década, la inteligencia artificial (IA) ha revolucionado numerosos campos, incluyendo la investigación científica. La IA se refiere a la simulación de procesos de inteligencia humana por parte de sistemas informáticos, que abarcan la adquisición de conocimiento, el razonamiento y la autocorrección. Por otro lado, la estadística inferencial es una rama de la estadística que permite hacer generalizaciones sobre una población a partir de una muestra, facilitando la toma de decisiones basadas en datos.

La convergencia de la inteligencia artificial y la estadística inferencial ha abierto nuevas vías para la investigación y la innovación. La capacidad de la IA para analizar grandes volúmenes de datos y reconocer patrones complejos se complementa con las herramientas de la estadística inferencial, que permiten validar hipótesis y estimar parámetros poblacionales (Rodríguez et al., 2025). Esta sinergia no solo mejora la precisión de los resultados, sino que también acelera el proceso de descubrimiento científico. La inteligencia artificial proporciona técnicas avanzadas que pueden manejar y extraer información útil de conjuntos de datos masivos, mientras que la estadística inferencial ofrece un marco riguroso para interpretar esos resultados y establecer conclusiones válidas.

La inteligencia artificial ha revolucionado la forma en que se aborda la investigación científica. A través de una variedad de métodos, la IA permite a los académicos analizar grandes volúmenes de datos, identificar patrones ocultos y realizar predicciones más precisas. El aprendizaje automático es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos que permiten a las computadoras aprender de los datos sin ser programadas explícitamente. Este enfoque se basa en la idea de que los sistemas pueden mejorar su rendimiento a medida que se exponen a más datos. En el contexto de la investigación científica, el aprendizaje automático se utiliza para clasificar datos, detectar anomalías y realizar predicciones.

Así como, en el ámbito de la biomedicina, los algoritmos de aprendizaje automático se emplean para analizar imágenes médicas y diagnosticar enfermedades con una precisión comparable a la de los expertos humanos. Además, en estudios epidemiológicos, estos algoritmos pueden ayudar a predecir brotes de enfermedades al analizar patrones en datos históricos.

Las redes neuronales son un componente fundamental del aprendizaje profundo, una subdisciplina del aprendizaje automático. Inspiradas en la estructura del cerebro humano, las redes neuronales consisten en capas de nodos (o neuronas) que procesan la información de manera jerárquica. Esta arquitectura permite a las redes neuronales manejar tareas complejas, como el reconocimiento de voz y la interpretación de imágenes.

En la investigación científica, las redes neuronales se han aplicado en diversas áreas, desde la predicción de propiedades químicas de nuevas moléculas hasta la generación de modelos climáticos. Su capacidad para aprender representaciones complejas de los datos las convierte en una herramienta poderosa para descubrir nuevas relaciones y patrones en conjuntos de datos masivos. El procesamiento del lenguaje natural (PLN) es un campo de la inteligencia artificial que se ocupa de la interacción entre las computadoras y el lenguaje humano. A través de técnicas de PLN, las máquinas pueden comprender, interpretar y generar texto en lenguaje natural, lo que facilita la automatización de tareas relacionadas con el análisis de texto (Taja, 2025).

En la investigación científica, el PLN se utiliza para analizar literatura académica, extraer información relevante de artículos y resúmenes, y realizar metaanálisis. Por ejemplo, mediante el uso de algoritmos de PLN, los académicos pueden identificar tendencias emergentes en un campo específico al analizar grandes volúmenes de publicaciones científicas. Esto no solo ahorra tiempo, sino que también puede revelar conexiones entre investigaciones que de otro modo podrían pasar desapercibidas.

Los métodos de inteligencia artificial, incluidos el aprendizaje automático, las redes neuronales y el procesamiento del lenguaje natural,

están transformando la investigación científica. Estas técnicas no solo mejoran la capacidad de los académicos para analizar datos complejos, sino que también abren nuevas posibilidades para la innovación y el descubrimiento en diversas disciplinas. La estadística inferencial consiente a los académicos en tomar decisiones y hacer predicciones sobre una población basándose en una muestra de datos. A través de diversas técnicas, los científicos pueden analizar la información recolectada y extraer conclusiones que van más allá de los datos observados.

Las pruebas de hipótesis son procedimientos estadísticos que permiten evaluar afirmaciones sobre parámetros poblacionales. Estas pruebas comienzan con la formulación de dos hipótesis: la hipótesis nula (H0), que representa una afirmación que se busca refutar, y la hipótesis alternativa (H1), que es lo que los académicos realmente quieren demostrar (Dagnino, 2014). Utilizando datos de una muestra, los académicos calculan un valor p que indica la probabilidad de observar los resultados si la hipótesis nula es cierta. Dependiendo del nivel de significancia preestablecido, se puede decidir rechazar o no la hipótesis nula. Este enfoque es esencial en campos como la medicina, donde se evalúa la efectividad de nuevos tratamientos.

Los intervalos de confianza son una herramienta que proporciona un rango estimado dentro del cual se espera que se encuentre un parámetro poblacional con un cierto nivel de confianza. Comúnmente, se utiliza un intervalo de confianza del 95%, lo que significa que hay un 95% de seguridad de que el verdadero valor del parámetro se encuentra dentro de dicho intervalo. Este método no solo permite a los académicos estimar un valor, sino que también ofrece una medida de la precisión de la estimación. En la investigación científica, los intervalos de confianza son útiles para comunicar

la incertidumbre asociada a los resultados, proporcionando una visión más completa de los hallazgos.

El análisis de regresión es una técnica que permite explorar las relaciones entre una variable dependiente y una o más variables independientes. A través de modelos de regresión, los académicos pueden predecir el comportamiento de la variable dependiente en función de los cambios en las variables independientes. Existen varios tipos de análisis de regresión, incluida la regresión lineal, que modela la relación lineal entre variables; y la regresión logística, que se utiliza para variables dependientes categóricas. Este análisis es particularmente valioso en estudios que requieren comprender cómo diferentes factores influyen en un resultado, como en investigaciones sobre factores de riesgo en salud o en estudios sociales.

Las técnicas de estadística inferencial son herramientas poderosas en la investigación científica, ya que permiten a los académicos hacer generalizaciones sobre poblaciones a partir de muestras y evaluar la validez de sus conclusiones. La adecuada aplicación de estas técnicas es crucial para garantizar que los hallazgos sean sólidos y fundamentados, lo que a su vez contribuye a la construcción del conocimiento científico. La integración de métodos de inteligencia artificial y estadística inferencial ha transformado la forma en que se lleva a cabo la investigación científica en diversas disciplinas.

En el ámbito médico, la combinación de IA y estadística inferencial ha permitido avances significativos en el diagnóstico y tratamiento de enfermedades. Se puede citar, el aprendizaje automático se utiliza para analizar grandes volúmenes de datos clínicos y genéticos, identificando patrones que pueden no ser evidentes a simple vista (Lanzagorta et al., 2022). Esto ha llevado a la creación de modelos predictivos que ayudan a los médicos

a anticipar la progresión de enfermedades, como el cáncer, y a personalizar tratamientos basados en las características individuales de cada paciente.

Además, las redes neuronales han demostrado ser particularmente efectivas en la interpretación de imágenes médicas, como resonancias magnéticas y radiografías. Estos sistemas pueden detectar anomalías con una precisión comparable a la de los radiólogos, lo que acelera el diagnóstico y mejora los resultados clínicos. Por otro lado, el procesamiento del lenguaje natural se utiliza para analizar literatura médica y registros clínicos, facilitando la extracción de información relevante y la identificación de tendencias en tratamientos y resultados.

La investigación ambiental se beneficia enormemente de la sinergia entre IA y estadística inferencial. Los modelos de predicción basados en IA pueden procesar datos de satélites, sensores y estaciones meteorológicas para evaluar el impacto del cambio climático, la contaminación y la pérdida de biodiversidad. En concreto, mediante el uso de algoritmos de aprendizaje automático, los científicos pueden predecir patrones de migración de especies y cambios en los ecosistemas, lo que permite implementar estrategias de conservación más efectivas. Las técnicas de estadística inferencial, como el análisis de regresión, son fundamentales para establecer relaciones causales entre variables ambientales y sus efectos en la salud pública. Esto permite a los académicos evaluar, por ejemplo, cómo la calidad del aire afecta la incidencia de enfermedades respiratorias en diferentes poblaciones, lo que resulta crucial para la formulación de políticas ambientales y de salud pública.

En el campo del desarrollo tecnológico, la IA y la estadística inferencial juegan un papel crucial en la innovación y mejora de productos y procesos. Las empresas utilizan algoritmos de aprendizaje automático para optimizar

la producción, predecir la demanda y mejorar la calidad de sus productos. La integración de técnicas estadísticas permite validar resultados y asegurar que las decisiones se basen en datos sólidos, minimizando riesgos y maximizando la eficiencia.

Así como, en la industria del software, las técnicas de procesamiento del lenguaje natural son utilizadas para desarrollar asistentes virtuales y chatbots que mejoran la interacción con los usuarios. Estos sistemas son capaces de aprender de las interacciones previas y ajustar sus respuestas, lo que se traduce en una experiencia más satisfactoria para el cliente. Asimismo, en el desarrollo de productos, el análisis de datos estadísticos ayuda a identificar tendencias de consumo y preferencias del mercado, guiando la dirección de la innovación.

Las aplicaciones de la inteligencia artificial y la estadística inferencial en la investigación científica son vastas y variadas, contribuyendo a avances significativos en la medicina, la protección del medio ambiente y el desarrollo tecnológico. Estas herramientas complementarias no solo optimizan los procesos de investigación, sino que también potencian la capacidad de los científicos para abordar desafíos complejos y mejorar la calidad de vida en la sociedad.

La integración de métodos de inteligencia artificial y estadística inferencial en la investigación científica representa un avance significativo en la forma en que abordamos y resolvemos problemas complejos. La combinación de estas dos disciplinas permite a los académicos extraer conocimientos más profundos y precisos de los datos, optimizando así el proceso de descubrimiento y validación de hipótesis.

Por un lado, la inteligencia artificial, a través de sus técnicas de aprendizaje automático y redes neuronales, ofrece herramientas potentes para el análisis de grandes volúmenes de datos. Estas técnicas son capaces de identificar patrones y relaciones que podrían pasar desapercibidos mediante métodos tradicionales. Por otro lado, la estadística inferencial proporciona un marco riguroso para realizar inferencias sobre poblaciones a partir de muestras, garantizando que los resultados obtenidos sean válidos y generalizables.

La sinergia entre estas metodologías no solo mejora la precisión de los resultados, sino que también permite a los académicos abordar preguntas más ambiciosas y complejas. En campos como la medicina, la combinación de algoritmos de inteligencia artificial con análisis estadísticos ha llevado a avances en diagnósticos y tratamientos personalizados. En el ámbito ambiental, esta integración ha facilitado la modelización de fenómenos complejos y la evaluación de políticas efectivas para la conservación. Además, la integración de métodos fomenta una cultura de colaboración interdisciplinaria, donde estadísticos, científicos de datos e académicos de diversas áreas trabajan juntos para abordar problemas de relevancia social. Esta colaboración no solo enriquece el proceso de investigación, sino que también promueve la innovación y la generación de soluciones más efectivas.

La unión de la inteligencia artificial y la estadística inferencial en la investigación científica es fundamental para el avance del conocimiento. Esta integración no solo amplifica las capacidades analíticas de los académicos, sino que también contribuye a un enfoque más holístico y riguroso en la búsqueda de respuestas a las preguntas más apremiantes de nuestra era. La continua evolución de estas metodologías sugiere que su impacto en la ciencia

seguirá creciendo, ofreciendo nuevas oportunidades para el descubrimiento y la innovación orientado a los datos.

Capítulo II

Comparación de Medias y Análisis de Datos Binarios en Mediciones Dependientes: Fundamentos, Técnicas y Aplicaciones Prácticas

En el ámbito de la investigación y el análisis de datos, la comparación de medias se ha convertido en una herramienta fundamental para entender las diferencias y similitudes entre grupos. Esta práctica es especialmente relevante cuando se trabaja con datos binarios, que representan una de las formas más simples de información, donde cada observación puede clasificarse en una de dos categorías. En particular, en estudios médicos, los resultados pueden ser positivos o negativos, aun cuando en investigaciones sociales, las respuestas pueden clasificarse como "sí" o "no".

La comparación de medias nos permite evaluar si las diferencias observadas entre grupos son significativas y no simplemente el resultado de variaciones aleatorias. En el contexto de datos binarios, aunque la medición de promedios puede parecer menos directa, existen métodos estadísticos que facilitan este análisis, permitiendo a los académicos extraer conclusiones valiosas a partir de conjuntos de datos que, en esencia, reflejan decisiones binarias. La importancia de esta comparación radica no solo en su aplicación en diversas disciplinas, como la medicina, la psicología y las ciencias sociales, sino también en su capacidad para informar decisiones críticas.

2.1 Fundamentos de la estadística en datos binarios

Los datos binarios son un tipo de variable categórica que solo puede tomar dos valores distintos, comúnmente representados como 0 y 1, o "sí" y "no". Este tipo de datos es ampliamente utilizado en diversas disciplinas, desde la investigación médica hasta la psicología y las ciencias sociales, debido a su capacidad para simplificar la representación de fenómenos complejos. A saber, en un estudio clínico, un paciente puede ser clasificado como "mejorado" o "no mejorado", lo que permite a los académicos analizar la efectividad de un tratamiento.

La importancia de los datos binarios radica en su omnipresencia en situaciones donde las decisiones deben basarse en resultados categóricos; es decir, la simplicidad de esta estructura permite que se utilicen métodos estadísticos específicos que pueden revelar patrones y relaciones que no serían evidentes en datos continuos, pues, el análisis de datos binarios se basa en varios principios estadísticos fundamentales (Villegas, 2019). Uno de los más importantes es la probabilidad, que permite a los académicos estimar la probabilidad de que ocurra un evento específico. En el contexto binario, esto se traduce en calcular la proporción de casos en los que se observa un resultado en relación con el total de casos analizados.

Otro concepto clave es la distribución binomial, que describe la probabilidad de obtener un número fijo de éxitos en una serie de ensayos independientes, cada uno con la misma probabilidad de éxito. Esta distribución es fundamental para realizar inferencias sobre los datos binarios y para calcular intervalos de confianza y pruebas de hipótesis. Existen diversas herramientas estadísticas que los académicos pueden utilizar para analizar datos binarios. Algunas de las más comunes incluyen:

i. *Pruebas de hipótesis:* Las pruebas como la prueba de chi-cuadrado permiten evaluar la relación entre dos variables categóricas y determinar si las diferencias observadas son estadísticamente significativas.

ii. *Regresión logística:* Esta técnica es esencial para modelar la relación entre una variable dependiente binaria y una o más variables independientes. A través de la regresión logística, los académicos pueden identificar factores que influyen en la probabilidad de que ocurra un evento.

iii. *Análisis de varianza (ANOVA):* Si bien tradicionalmente se asocia con datos continuos, el ANOVA se puede adaptar para analizar datos binarios mediante la comparación de proporciones.

iv. *Intervalos de confianza:* Calcular intervalos de confianza para proporciones permite a los académicos estimar el rango dentro del cual se espera que se encuentre la verdadera proporción en la población, lo que es crucial para la interpretación de resultados.

La comparación de medias en mediciones dependientes es un aspecto crucial en el análisis estadístico, especialmente cuando se trabaja con datos que provienen de las mismas unidades de observación en diferentes momentos o condiciones. Este enfoque permite a los académicos evaluar el impacto de un tratamiento o intervención sobre una misma muestra, eliminando la variabilidad individual que podría influir en los resultados y proporcionando una comprensión más clara de los efectos.

Las mediciones dependientes se refieren a situaciones en las que se toman múltiples observaciones de la misma entidad o individuo, esto puede incluir, a saber, mediciones antes y después de un tratamiento, o evaluaciones en diferentes momentos de una intervención. Al utilizar este diseño, los

académicos pueden controlar la variabilidad que podría surgir de diferencias individuales, lo que permite una comparación más precisa y significativa de las medias.

Existen varias técnicas estadísticas que se pueden utilizar para comparar medias en muestras dependientes, entre las más comunes se encuentra la prueba t de Student para muestras relacionadas, que se utiliza cuando se desea comparar las medias de dos grupos relacionados. Esta prueba evalúa si hay una diferencia estadísticamente significativa entre las dos medias y se basa en la diferencia de las puntuaciones de cada par de observaciones. Otra técnica relevante es el análisis de varianza (ANOVA) para medidas repetidas, que se aplica cuando se tienen más de dos condiciones o momentos de medición (Maestro et al., 2017). Este enfoque permite evaluar si hay diferencias significativas en las medias entre las diferentes condiciones, teniendo en cuenta la correlación entre las mediciones.

Además, se pueden emplear métodos no paramétricos, como la prueba de Wilcoxon para muestras relacionadas, en casos donde los datos no cumplen con los supuestos de normalidad requeridos por las pruebas paramétricas. Estas técnicas son esenciales para garantizar que los resultados sean robustos y confiables. La interpretación de los resultados obtenidos de las comparaciones de medias en mediciones dependientes debe hacerse con cautela. Es fundamental considerar no solo el valor p, que indica la significancia estadística, sino también el tamaño del efecto, que proporciona información sobre la magnitud de la diferencia observada. Un tamaño de efecto grande puede indicar que la intervención tiene un impacto práctico significativo, a la vez que un valor p bajo sin un tamaño de efecto adecuado puede no ser relevante en la práctica.

Además, es esencial tener en cuenta el contexto del estudio y la relevancia clínica de los resultados. La interpretación debe considerar no solo la estadística, sino también la aplicabilidad de los hallazgos en situaciones del mundo real. Esto es especialmente importante en campos como la medicina y la psicología, donde las decisiones basadas en los resultados del análisis pueden tener implicaciones significativas para la salud y el bienestar de las personas.

La comparación de medias en mediciones dependientes es un área rica y compleja dentro del análisis estadístico que permite a los académicos obtener conclusiones más precisas al estudiar el efecto de intervenciones sobre los mismos sujetos. La correcta aplicación y análisis de estas técnicas son esenciales para avanzar en la investigación en diversas disciplinas.

En el ámbito de la medicina, la comparación de medias y el análisis de datos binarios son esenciales para evaluar la eficacia de tratamientos y procedimientos; en concreto, en un estudio que investiga la efectividad de un nuevo fármaco para el tratamiento de la hipertensión, se pueden recoger datos sobre la presión arterial de los pacientes antes y después de la administración del medicamento. Aquí, las mediciones son dependientes, ya que se están comparando las mismas personas en dos momentos distintos. Utilizando pruebas estadísticas como la prueba t para muestras relacionadas, se puede determinar si hay una diferencia significativa en las medias de las presiones arteriales, lo que podría indicar la efectividad del tratamiento.

Además, los datos binarios pueden ser utilizados para evaluar la respuesta de los pacientes al tratamiento, es decir, si lograron o no alcanzar un nivel de presión arterial deseado. Este tipo de análisis permite a los académicos tomar decisiones informadas sobre la implementación de nuevos tratamientos

basados en evidencia sólida. En el campo de la psicología, la comparación de medias en mediciones dependientes es igualmente crucial. Así como, un estudio podría investigar el impacto de una intervención psicológica en los niveles de ansiedad de los participantes. Se pueden medir los niveles de ansiedad antes y después de la intervención, permitiendo así a los académicos analizar si la intervención ha tenido un efecto significativo.

Por otro lado, en ciencias sociales, los datos binarios son frecuentemente utilizados para analizar comportamientos y actitudes. Un ejemplo podría ser un estudio sobre la aceptación de un programa educativo en una comunidad, donde los encuestados responden "sí" o "no" a preguntas sobre su interés en participar. Al comparar las proporciones de respuestas afirmativas antes y después de la implementación del programa, se pueden extraer conclusiones sobre la efectividad y aceptación del mismo.

Los hallazgos derivados de la comparación de medias y el análisis de datos binarios tienen profundas implicaciones en la toma de decisiones en diversos sectores. En el ámbito médico, los resultados pueden influir en la adopción de nuevos tratamientos; en psicología, pueden guiar la implementación de intervenciones terapéuticas efectivas; y en ciencias sociales, pueden ayudar a diseñar programas que respondan mejor a las necesidades de la comunidad.

Las aplicaciones prácticas de la comparación de medias observadas y el análisis de datos binarios son vastas y variadas, abarcando desde la investigación médica hasta la psicología y las ciencias sociales. Estos enfoques no solo permiten a los académicos obtener conclusiones significativas, sino que también informan decisiones críticas que pueden mejorar la calidad de vida y el bienestar de las personas.

La capacidad de analizar y comparar datos binarios no solo permite a los académicos identificar patrones y tendencias significativas, sino que también facilita la toma de decisiones informadas. En campos como la medicina, donde las decisiones pueden tener implicaciones de vida o muerte, la correcta interpretación de estos datos es crucial. Asimismo, en psicología y ciencias sociales, donde las respuestas pueden estar condicionadas por múltiples factores, el análisis de datos binarios ofrece una perspectiva que puede desvelar dinámicas complejas en el comportamiento humano.

Mirando hacia el futuro, es evidente que el análisis de datos binarios seguirá evolucionando. Con el auge de la inteligencia artificial y el aprendizaje automático, se prevé que las técnicas de análisis se vuelvan más sofisticadas, permitiendo una interpretación más profunda y precisa de los datos. Además, la integración de métodos estadísticos avanzados con plataformas de análisis de datos en tiempo real permitirá a los académicos manejar volúmenes de datos cada vez mayores y más complejos, lo que abrirá nuevas oportunidades para la investigación.

Por otro lado, el desarrollo de software y herramientas estadísticas más accesibles está democratizando el acceso a técnicas de análisis avanzadas, lo que significa que un número creciente de académicos, independientemente de su formación estadística, podrá realizar análisis significativos y contribuir al conocimiento en sus respectivas áreas. El análisis de datos binarios es un campo en constante crecimiento que promete revolucionar la forma en que interpretamos e interactuamos con la información.

2.2 Prueba de Equivalencia (TOST) en Aprendizaje Automático: Fundamentos, Aplicaciones y Desafíos

La Prueba de Equivalencia, específicamente el enfoque conocido como Two One-Sided Tests (TOST), se presenta como una herramienta estadística crucial para abordar esta necesidad. A diferencia de las pruebas tradicionales que buscan demostrar una diferencia significativa entre grupos, TOST se enfoca en verificar que la diferencia entre dos estimaciones se mantenga dentro de límites específicos, es decir, que no sea significativamente diferente de cero en un sentido práctico. Esto es particularmente relevante en el contexto del aprendizaje automático, donde se busca comprobar que un nuevo modelo no solo es diferente, sino que es comparable o igual a uno existente (Paixão et al., 2022).

La validación de modelos en aprendizaje automático es un proceso crítico que puede influir en la adopción y la confianza en estos sistemas. La Prueba de Equivalencia proporciona una metodología robusta para hacer esta validación, permitiendo a los académicos y profesionales del sector demostrar que un modelo nuevo no solo es eficaz, sino que es equivalente a un modelo establecido. Esto es especialmente útil en situaciones donde los recursos son limitados y se deben seleccionar entre múltiples enfoques. La capacidad de demostrar equivalencia puede facilitar la toma de decisiones informadas y minimizar riesgos al implementar nuevas tecnologías.

La Prueba de Equivalencia, comúnmente conocida por su acrónimo TOST (Two One-Sided Tests), es una herramienta estadística crucial en el contexto del aprendizaje automático que permite establecer si dos métodos o modelos son equivalentes en términos de desempeño. La Prueba de Equivalencia se basa en la premisa de que, en lugar de buscar diferencias

significativas entre dos grupos, el objetivo es demostrar que estas diferencias se mantienen dentro de un rango predefinido de equivalencia. Esto es particularmente relevante en el aprendizaje automático, donde se busca validar que un nuevo modelo no es significativamente inferior a un modelo existente, o viceversa.

El enfoque TOST implica la realización de dos pruebas de hipótesis unilaterales. La primera prueba evalúa si el nuevo modelo es significativamente mejor que el modelo de referencia, entre tanto que la segunda prueba verifica si el nuevo modelo no es significativamente peor. Si ambas hipótesis se aceptan, se puede concluir que los modelos son equivalentes dentro de los límites establecidos.

Para aplicar la Prueba de Equivalencia, es esencial definir criterios claros y cuantificables que determinen el rango de equivalencia. Estos criterios suelen expresarse en términos de una diferencia aceptable en métricas de rendimiento, como la precisión, la sensibilidad o el área bajo la curva (AUC). Para ilustrar, en un estudio de validación de modelos, se podría establecer que la diferencia en precisión entre el nuevo modelo y el modelo de referencia no debe exceder el 5%.

La elección de estos criterios es crítica y debe basarse en el contexto específico del problema, así como en las expectativas de los stakeholders. Un criterio demasiado amplio puede conducir a la aceptación de modelos que en realidad no son comparables, aun cuando que un criterio demasiado estricto puede resultar en la desestimación de modelos que podrían ser útiles. La implementación de la Prueba de Equivalencia requiere el uso de diversos métodos estadísticos que permiten calcular y evaluar las diferencias de

rendimiento entre los modelos. Algunos de los métodos más comunes incluyen:

i. *Intervalos de confianza:* Se utilizan para estimar el rango de diferencias de rendimiento entre los dos modelos. Si el intervalo de confianza de la diferencia se encuentra dentro de los límites de equivalencia, se puede concluir que los modelos son equivalentes.
ii. *Pruebas t de Student:* Se aplican para comparar las medias de dos grupos y determinar si hay diferencias significativas en su rendimiento. En el contexto de TOST, se realizarían dos pruebas t de una sola cola.
iii. *Análisis de varianza (ANOVA):* Si bien no es tan común en la TOST, en algunos casos se puede utilizar para evaluar si hay diferencias significativas entre más de dos grupos.
iv. *Simulaciones y bootstrap:* Estas técnicas permiten la estimación de la variabilidad de las métricas de rendimiento y son útiles para construir intervalos de confianza robustos. Los fundamentos de la Prueba de Equivalencia son esenciales para asegurar que los modelos en el aprendizaje automático sean evaluados de manera justa y rigurosa.

El uso de la Prueba de Equivalencia (TOST) en el ámbito del aprendizaje automático ha ganado relevancia en los últimos años, dado que permite establecer comparaciones significativas entre modelos y algoritmos, asegurando que las diferencias en rendimiento sean estadísticamente irrelevantes. Una de las aplicaciones más comunes de TOST en el aprendizaje automático es la validación de modelos (Paixão et al., 2022). Cuando se desarrolla un nuevo modelo de aprendizaje automático, es crucial evaluar su desempeño en comparación con un modelo existente o un estándar de

referencia. La Prueba de Equivalencia permite a los académicos y profesionales determinar si el nuevo modelo es estadísticamente equivalente al modelo de referencia en términos de métricas de rendimiento, como la precisión, la recuperación o el F1-score.

Este enfoque es especialmente útil en contextos donde la interpretación de los resultados es crítica. Para ilustrar, en aplicaciones médicas, donde un modelo de diagnóstico debe ser tan efectivo como un método estándar, TOST proporciona una forma robusta de validar que el nuevo modelo no solo es bueno, sino que funciona igual de bien que el método existente. La comparación de diferentes algoritmos de aprendizaje automático es otra área donde TOST demuestra su utilidad. En la práctica, los académicos a menudo experimentan con múltiples algoritmos para resolver un problema específico. Sin embargo, no siempre es obvio si un algoritmo es realmente mejor que otro. TOST permite establecer si las diferencias observadas en el rendimiento de los algoritmos son estadísticamente significativas o si están dentro de un rango aceptable de equivalencia.

En particular, al comparar un algoritmo de árbol de decisión con un algoritmo de redes neuronales, TOST puede ayudar a determinar si las mejoras en la precisión del modelo de redes neuronales justifican su mayor complejidad computacional o si el modelo de árbol de decisión es una alternativa suficientemente buena.

Existen varios estudios de caso que ilustran la aplicación de TOST en el aprendizaje automático. Un ejemplo notable se encuentra en el ámbito del procesamiento del lenguaje natural (NLP). En un estudio reciente, los académicos utilizaron TOST para comparar un nuevo modelo de traducción automática con un sistema de traducción establecido. A través de la prueba,

pudieron demostrar que, si bien el nuevo modelo tenía ligeras variaciones en la calidad de las traducciones, estas caían dentro del rango de equivalencia predefinido, lo que justificó su implementación en aplicaciones comerciales.

Otro caso se da en la clasificación de imágenes, donde los desarrolladores de un nuevo modelo de red neuronal convolucional (CNN) utilizaron TOST para validar que su modelo proporcionaba resultados equivalentes a un modelo de referencia ampliamente utilizado. Esto les permitió publicar sus hallazgos con confianza, sabiendo que su modelo no solo era competitivo sino que también cumplía con los estándares de equivalencia. La Prueba de Equivalencia se ha convertido en una herramienta invaluable en el aprendizaje automático, facilitando la validación de modelos, la comparación de algoritmos y proporcionando un marco sólido para la interpretación de resultados en investigaciones.

A pesar de los beneficios que ofrece la Prueba de Equivalencia (TOST) en el contexto del aprendizaje automático, esta metodología no está exenta de limitaciones. En primer lugar, la elección de los márgenes de equivalencia puede ser un desafío significativo. La definición de un intervalo que capture adecuadamente la equivalencia entre modelos puede ser subjetiva y, en consecuencia, influir en las conclusiones. Además, la TOST asume que los datos son independientes y que siguen una distribución normal, lo cual no siempre se cumple en escenarios del mundo real. Esto puede llevar a resultados engañosos si los supuestos subyacentes no se verifican adecuadamente.

La interpretación de los resultados de una prueba de equivalencia requiere una comprensión cuidadosa de los datos y del contexto en el que se aplican. Aunque una prueba puede indicar que dos modelos son equivalentes dentro

de un margen específico, esto no necesariamente implica que ambos modelos sean igualmente adecuados para una tarea particular. Los resultados deben ser contextualizados, considerando otros aspectos como la complejidad del modelo, el tiempo de entrenamiento y la interpretabilidad. Además, los resultados de la TOST deben complementarse con otros métodos de evaluación para proporcionar una visión más completa del rendimiento del modelo.

Para abordar las limitaciones y mejorar la aplicabilidad de la Prueba de Equivalencia en el aprendizaje automático, se recomienda que los futuros estudios consideren varios enfoques. En primer lugar, es crucial realizar un análisis de sensibilidad que explore cómo los cambios en los márgenes de equivalencia afectan los resultados. Asimismo, se sugiere combinar la TOST con otros métodos estadísticos y métricas de rendimiento, como la validación cruzada y las curvas ROC, para obtener una evaluación más robusta. Por último, fomentar el diálogo entre estadísticos y expertos en aprendizaje automático puede resultar en el desarrollo de metodologías más efectivas y adaptadas a las particularidades de los modelos en este campo en constante evolución.

Aunque la Prueba de Equivalencia ofrece un marco valioso para evaluar la equivalencia de modelos, es fundamental tener en cuenta sus limitaciones y adoptar un enfoque integral que permita una evaluación más precisa y contextualizada en el ámbito del aprendizaje automático. La Prueba de Equivalencia, y en particular el enfoque de TOST (Two One-Sided Tests), se ha revelado como una herramienta valiosa en el contexto del aprendizaje automático (Paixão et al., 2022). Los conceptos fundamentales de TOST nos permiten establecer criterios claros para evaluar la equivalencia entre diferentes métodos, lo que es esencial en un campo donde la selección de

modelos puede influir drásticamente en los resultados. Además, hemos examinado cómo TOST se aplica en estudios de caso concretos, demostrando su utilidad en la práctica.

La implementación de la Prueba de Equivalencia en el aprendizaje automático no solo ayuda a validar modelos de manera más efectiva, sino que también fomenta una cultura de rigor científico y transparencia en la evaluación de algoritmos. Esto tiene implicaciones significativas no solo para el desarrollo de tecnologías más robustas, sino también para la confianza del usuario en los sistemas de inteligencia artificial (Najem et al., 2025).

A pesar de las ventajas que ofrece TOST, es vital reconocer sus limitaciones y los desafíos asociados con su aplicación. Para avanzar en este campo, futuras investigaciones deberían centrarse en la adaptación de métodos de equivalencia a contextos más complejos y dinámicos, así como en el desarrollo de nuevas métricas que complementen las pruebas tradicionales. También sería beneficioso explorar la integración de TOST con enfoques más modernos de aprendizaje automático, como el aprendizaje profundo, donde la validación de modelos se enfrenta a nuevos retos. En última instancia, el camino hacia adelante debe incluir una colaboración interdisciplinaria que combine estadísticas, teoría del aprendizaje y aplicación práctica para maximizar el impacto de la Prueba de Equivalencia en el aprendizaje automático.

2.3 Comparativa de la Prueba de Kruskal-Wallis y la Prueba de Mood en el Contexto del Aprendizaje Automático

En el campo del aprendizaje automático, la evaluación y comparación de algoritmos son fundamentales para garantizar que se seleccionen las mejores técnicas para resolver problemas específicos. Entre las diversas

pruebas estadísticas disponibles, la prueba de Kruskal-Wallis y la prueba de Mood se destacan por su capacidad para realizar comparaciones entre grupos no paramétricos, es decir, aquellos que no siguen una distribución normal.

La prueba de Kruskal-Wallis, una extensión de la prueba de rangos de Wilcoxon, se utiliza comúnmente para evaluar si hay diferencias significativas en las medianas de tres o más grupos independientes. Por su parte, la prueba de Mood, también conocida como prueba de la mediana, se centra en comparar las medianas de dos o más grupos, proporcionando una alternativa robusta a los métodos paramétricos tradicionales. La prueba de Kruskal-Wallis es una técnica estadística no paramétrica que se utiliza para determinar si existen diferencias significativas entre las medianas de tres o más grupos independientes. Esta prueba es particularmente útil cuando los datos no cumplen con los supuestos necesarios para realizar un análisis de varianza (ANOVA) tradicional, como la normalidad de las distribuciones y la homogeneidad de las varianzas.

La prueba de Kruskal-Wallis, desarrollada por William Kruskal y W. Allen Wallis en 1952, se basa en el orden de los datos en lugar de sus valores absolutos. Esto significa que, en lugar de evaluar directamente las medias, se clasifican todos los datos y se asignan rangos. La prueba evalúa si la distribución de los rangos es la misma en todos los grupos. Si se encuentran diferencias significativas, se puede inferir que al menos uno de los grupos difiere en su mediana respecto a los otros. El estadístico de la prueba, denominado H, se calcula utilizando la suma de los rangos de cada grupo y el tamaño de cada grupo. Un valor alto de H indica que hay diferencias significativas en las medianas. La significancia se determina comparando el valor calculado de H con una distribución chi-cuadrado con k-1 grados de libertad, donde k es el número de grupos.

En el ámbito del aprendizaje automático, la prueba de Kruskal-Wallis se aplica frecuentemente para comparar el rendimiento de diferentes algoritmos o modelos en función de diversas métricas de evaluación, como la precisión, la recuperación o la F1-score. A saber, al evaluar múltiples modelos en un conjunto de datos, se puede utilizar esta prueba para determinar si hay diferencias significativas en la efectividad de los modelos, lo que puede guiar a los académicos a elegir el enfoque más prometedor (Garrocho et al., 2024). Además, la prueba se puede utilizar en la selección de características, donde se busca identificar qué variables tienen un impacto significativo en la variable objetivo. Al evaluar las medianas de las características categóricas en relación con la variable dependiente, se pueden tomar decisiones informadas sobre qué atributos incluir en un modelo.

A pesar de su utilidad, la prueba de Kruskal-Wallis tiene ciertas limitaciones. En primer lugar, si bien es robusta frente a violaciones de normalidad, sigue siendo sensible a la presencia de valores atípicos, que pueden influir en los rangos y, por ende, en el resultado de la prueba. Además, la prueba solo indica si hay diferencias significativas entre las medianas de los grupos, pero no proporciona información sobre cuáles grupos son diferentes entre sí. Por lo tanto, en caso de que se rechace la hipótesis nula, se requiere realizar pruebas post hoc para identificar las diferencias específicas.

Otra consideración importante es que la prueba de Kruskal-Wallis asume que las distribuciones de los grupos son similares en forma. Si las distribuciones son muy diferentes, los resultados pueden ser engañosos. Por tanto, es esencial realizar un análisis exploratorio de los datos antes de aplicar esta prueba. La prueba de Kruskal-Wallis es una herramienta poderosa en el aprendizaje automático para comparar múltiples grupos, pero debe ser

utilizada con precaución y en el contexto apropiado, complementada con análisis adicionales cuando sea necesario.

La prueba de Mood, también conocida como la prueba de la mediana de Mood, es un procedimiento estadístico no paramétrico que se utiliza para evaluar si existen diferencias significativas entre las medianas de dos o más grupos independientes. A diferencia de la prueba de Kruskal-Wallis, que se centra en las distribuciones completas de los grupos, la prueba de Mood se enfoca específicamente en la mediana, lo que la convierte en una opción útil cuando las distribuciones de los grupos pueden no ser simétricas o pueden contener valores atípicos.

La prueba de Mood se basa en el cálculo de las diferencias entre cada par de grupos en términos de sus medianas. En esencia, la prueba evalúa si las diferencias observadas en las medianas son lo suficientemente grandes como para ser consideradas estadísticamente significativas, es decir, si podrían haber ocurrido por azar. Este enfoque la convierte en una herramienta valiosa en el análisis de datos en contextos donde la normalidad de la distribución no puede ser asumida.

En el contexto del aprendizaje automático, la prueba de Mood se utiliza frecuentemente para comparar el rendimiento de diferentes algoritmos o modelos. Así como, al evaluar la precisión de distintos modelos de clasificación en un conjunto de datos, los académicos pueden utilizar la prueba de Mood para determinar si las diferencias en las medianas de las métricas de rendimiento (como la tasa de aciertos) son significativas. Esto es especialmente relevante en situaciones donde los resultados pueden estar afectados por la variabilidad inherente de los datos o la presencia de variaciones que podrían distorsionar las métricas de rendimiento promedio.

La prueba de Mood permite a los científicos de datos y a los ingenieros de aprendizaje automático tomar decisiones informadas sobre qué modelos son superiores en función de sus rendimientos empíricos, proporcionando una base estadística sólida para seleccionar el modelo más adecuado para una tarea específica. Aunque tanto la prueba de Mood como la prueba de Kruskal-Wallis son herramientas no paramétricas que se utilizan para comparar grupos, existen diferencias clave entre ambas. La prueba de Kruskal-Wallis evalúa si hay diferencias en las distribuciones completas de los grupos, mientras que la prueba de Mood se centra exclusivamente en las diferencias en las medianas. Esto significa que la prueba de Mood puede ser más sensible a cambios en la mediana, especialmente en distribuciones asimétricas o en presencia de variaciones.

Además, la prueba de Mood es generalmente más adecuada cuando se comparan solo dos grupos, aunque también se puede extender a más de dos. En contraste, la prueba de Kruskal-Wallis es más eficiente y robusta cuando se trata de comparar tres o más grupos. La elección entre la prueba de Mood y la prueba de Kruskal-Wallis dependerá del contexto específico del análisis y de los supuestos sobre la distribución de los datos. Ambas pruebas ofrecen valiosas perspectivas en el análisis de datos en el aprendizaje automático, pero su aplicabilidad puede variar según la naturaleza de los datos y los objetivos del estudio.

La comparación entre la prueba de Kruskal-Wallis y la prueba de Mood es fundamental para entender cuál de estas herramientas estadísticas puede ser más adecuada en diferentes contextos dentro del aprendizaje automático. Ambas pruebas son no paramétricas y se utilizan para evaluar diferencias entre grupos, pero tienen características y aplicaciones particulares que las distinguen (Garrocho et al., 2024).

La prueba de Kruskal-Wallis es generalmente utilizada cuando se desea comparar tres o más grupos independientes. Es especialmente útil en situaciones donde se tiene una variable dependiente ordinal o continua y se quiere evaluar si hay diferencias significativas en la mediana de las distribuciones de los grupos. Esto es, en un estudio donde se comparan tres algoritmos de clasificación en términos de precisión, Kruskal-Wallis puede ayudar a determinar si al menos uno de los algoritmos presenta un rendimiento significativamente diferente.

Por otro lado, la prueba de Mood es más apropiada cuando se compara la variabilidad entre dos o más grupos. Esta prueba es particularmente valiosa en el contexto de algoritmos de aprendizaje automático que requieren una evaluación de la estabilidad o robustez de su rendimiento. En concreto, en un escenario donde se desea analizar la variabilidad en los errores de predicción de dos modelos diferentes, la prueba de Mood puede proporcionar información sobre si las diferencias en la variabilidad son significativas.

Una de las principales ventajas de la prueba de Kruskal-Wallis es su capacidad para manejar conjuntos de datos que no cumplen con los supuestos de normalidad y homogeneidad de varianzas, lo que la hace más flexible en la práctica. Sin embargo, su desventaja radica en que solo indica que hay diferencias significativas entre los grupos, pero no especifica cuáles son esos grupos que difieren.

La prueba de Mood, si bien también es no paramétrica, tiene la ventaja de enfocarse en la variabilidad, lo cual es crucial en el aprendizaje automático, donde la estabilidad de un modelo puede ser tan importante como su precisión. Sin embargo, al igual que Kruskal-Wallis, esta prueba requiere que

los datos sean independientes y puede ser menos conocida y utilizada, lo que podría limitar su aplicación en algunos contextos.

Un ejemplo práctico del uso de la prueba de Kruskal-Wallis en aprendizaje automático podría ser la evaluación de diferentes técnicas de preprocesamiento de datos. Supongamos que se está trabajando con un conjunto de datos en el que se aplican diferentes métodos de normalización y se quiere determinar si alguno de estos métodos resulta en un rendimiento significativamente mejor en un modelo de clasificación. En este caso, la prueba de Kruskal-Wallis podría ayudar a identificar si hay diferencias en el rendimiento entre las técnicas.

En contraste, un ejemplo de la aplicación de la prueba de Mood podría involucrar la comparación de la variabilidad en las tasas de error de dos algoritmos de regresión bajo diferentes condiciones de ruido en los datos. Aquí, la prueba de Mood podría ser utilizada para evaluar si uno de los algoritmos es más robusto frente a las fluctuaciones en los datos que el otro, proporcionando información valiosa para la selección del modelo más adecuado.

Tanto la prueba de Kruskal-Wallis como la prueba de Mood ofrecen enfoques complementarios para la evaluación de algoritmos en el aprendizaje automático, cada una con sus propias fortalezas y limitaciones. La elección entre una u otra dependerá del objetivo específico del análisis y de la naturaleza de los datos en cuestión. En el ámbito del aprendizaje automático, la evaluación y comparación de algoritmos son esenciales para garantizar que los modelos desarrollados sean efectivos y precisos. Las pruebas estadísticas, como la prueba de Kruskal-Wallis y la prueba de Mood, desempeñan un papel

crucial en este proceso, proporcionando herramientas robustas para analizar diferencias en el rendimiento de múltiples grupos o condiciones.

La prueba de Kruskal-Wallis se destaca por su capacidad para comparar más de dos grupos independientes y es particularmente útil cuando los supuestos de normalidad no se cumplen. Su aplicación en el aprendizaje automático permite a los académicos validar la efectividad de diferentes algoritmos o configuraciones mediante un análisis no paramétrico, lo que la convierte en una opción valiosa en contextos donde los datos pueden ser ordinales o no distribuidos normalmente.

Por otro lado, la prueba de Mood, con su enfoque en la comparación de varianzas, ofrece una perspectiva complementaria. Al evaluar la homogeneidad de las varianzas entre grupos, esta prueba puede ser esencial para detectar inconsistencias que podrían afectar la interpretación de los resultados de los algoritmos. Aunque menos conocida que la prueba de Kruskal-Wallis, su uso en la evaluación de algoritmos permite a los académicos profundizar en la estabilidad y confiabilidad de sus modelos.

Al elegir entre estas pruebas, es fundamental considerar el tipo de datos y los objetivos del análisis. En última instancia, el uso adecuado de estas herramientas estadísticas no solo mejora la calidad de las investigaciones en aprendizaje automático, sino que también contribuye a la producción de modelos más robustos y confiables. La integración de la prueba de Kruskal-Wallis y la prueba de Mood en el análisis de algoritmos de aprendizaje automático ofrece un enfoque más completo para la evaluación de rendimiento.

Capítulo III

Comparación de Series de Datos Binarios: Varianzas y Proporciones en Análisis Estadístico

En la era del big data y la analítica avanzada, la capacidad de analizar y comparar datos se ha vuelto esencial para diversas disciplinas, desde la investigación científica hasta el marketing. Una de las formas más sencillas y efectivas de representar información es a través de datos binarios, que se caracterizan por tener solo dos posibles valores: generalmente 0 y 1, o "sí" y "no". Esta simplicidad permite una rápida interpretación y análisis, lo que hace que los datos binarios sean particularmente útiles en una variedad de contextos, incluyendo encuestas, experimentos y estudios de comportamiento (Najem et al., 2025).

Los datos binarios son aquellos que solo pueden tomar uno de dos valores posibles. Pongamos el caso de, en un estudio donde se pregunta a los participantes si prefieren un producto A o un producto B, la respuesta puede codificarse como 1 para "sí" (preferencia por el producto A) y 0 para "no" (preferencia por el producto B). Esta codificación simplifica el análisis estadístico y permite la utilización de diversas técnicas cuantitativas para extraer conclusiones significativas.

La comparación de series de datos binarios es crucial en muchas áreas de investigación y práctica. Permite a los académicos y analistas identificar patrones, tendencias y diferencias significativas entre grupos o condiciones. Por ejemplo, en estudios clínicos, comparar la tasa de éxito de un tratamiento

en dos grupos de pacientes (tratamiento versus control) puede proporcionar información valiosa sobre la efectividad del tratamiento. Además, en el ámbito del marketing, comparar tasas de conversión entre campañas puede ayudar a optimizar estrategias y maximizar el retorno de la inversión.

La varianza es una medida estadística que cuantifica la dispersión de un conjunto de datos respecto a su media. En el contexto de datos binarios, que solo pueden tomar los valores 0 y 1 (a saber, éxito o fracaso, presencia o ausencia), la varianza proporciona información sobre la variabilidad de los resultados de una serie de observaciones. A la vez que en datos continuos la varianza refleja la extensión de los valores en una escala numérica, en datos binarios, la varianza se interpreta como la medida de la incertidumbre o la diversidad en la ocurrencia de eventos.

El cálculo de la varianza para una serie de datos binarios es relativamente sencillo. Primero, se debe calcular la proporción de éxitos (1s) en la serie de datos, representada como p. A partir de esta proporción, la varianza se puede calcular utilizando la fórmula:

$$
\text{Varianza} = p(1 - p)
$$

Aquí, p es la probabilidad de éxito y $(1 - p)$ es la probabilidad de fracaso. Esta fórmula es especialmente útil dado que, debido a la naturaleza binaria de los datos, la varianza dependerá directamente de la proporción de éxitos. Es importante destacar que la varianza alcanzará su valor máximo cuando $p = 0.5$, lo que indica la mayor incertidumbre en los resultados.

La interpretación de la varianza en datos binarios permite a los académicos entender no solo la cantidad de éxitos y fracasos, sino también la estabilidad o la inestabilidad de estos resultados. Una varianza alta sugiere que hay una gran diversidad en los resultados, lo que podría indicar que hay factores subyacentes que influyen en la ocurrencia de eventos. Por el contrario, una varianza baja indica que los resultados son más predecibles y consistentes.

Esto es, si estamos analizando la eficacia de un nuevo tratamiento médico y encontramos una varianza baja en los resultados binarios (éxito o fracaso del tratamiento), esto puede sugerir que el tratamiento tiene un efecto consistente entre los pacientes. En cambio, si la varianza es alta, podría ser necesario investigar más a fondo las diferencias individuales que afectan la respuesta al tratamiento. La varianza en datos binarios no solo proporciona una medida de dispersión, sino que también ofrece teorías sobre la confiabilidad y la consistencia de los resultados, lo que es esencial para la toma de decisiones en múltiples campos, desde la investigación científica hasta el análisis de negocios.

3.1 Proporciones en datos binarios

La proporción es una medida que expresa la relación entre una parte y el todo, y en el contexto de datos binarios, se refiere a la frecuencia relativa de un evento específico frente a la totalidad de observaciones. En datos binarios, donde las respuestas son generalmente codificadas como 0 (no éxito) y 1 (éxito), la proporción de éxitos se puede calcular como el número de casos de éxito dividido por el número total de casos. Se puede citar, si estamos analizando la efectividad de un tratamiento médico a través de un ensayo

clínico y encontramos que 80 de 100 pacientes presentan una mejora, la proporción de éxito sería 0.8 o 80%.

El cálculo de proporciones en datos binarios es fundamental para evaluar y comparar el desempeño de diferentes grupos o condiciones. Para calcular la proporción, se utiliza la fórmula:

\[

P = \frac{X}{N}

\]

donde \(P \) es la proporción, \(X \) es el número de éxitos y \(N \) es el número total de observaciones. Esta medida es clave en la investigación, ya que permite a los académicos entender la prevalencia de un evento y facilita la comparación entre grupos. En concreto, en estudios de salud pública, comparar la proporción de individuos que contraen una enfermedad en dos poblaciones diferentes puede revelar patrones importantes que guíen políticas de prevención.

Comparar proporciones entre diferentes series de datos binarios es esencial para identificar diferencias significativas. Para realizar esta comparación, se pueden utilizar diversas técnicas estadísticas, incluyendo pruebas de hipótesis como la prueba de proporciones. Esta prueba permite determinar si la diferencia observada entre dos proporciones es estadísticamente significativa o si podría ser el resultado de variabilidad aleatoria.

Para ilustrar, si en un estudio se observa que la proporción de éxito de un nuevo tratamiento es del 70% en un grupo de pacientes y del 50% en otro grupo, un análisis de comparación de proporciones puede ayudar a establecer

si este 20% de diferencia es significativo. Este tipo de análisis no solo proporciona información sobre la efectividad de las intervenciones, sino que también puede influir en la toma de decisiones clínicas y en el diseño de futuras investigaciones. Las proporciones en datos binarios son una herramienta valiosa para la interpretación de resultados en diversas disciplinas. Su cálculo y comparación permiten a los académicos y profesionales evaluar la efectividad de intervenciones, identificar tendencias y formular recomendaciones basadas en evidencia.

Por otro lado, al abordar las proporciones, hemos definido su concepto y mostrado cómo calcularlas en series de datos binarios. La comparación de proporciones entre diferentes series permite identificar diferencias significativas y evaluar la relación entre variables, lo que es crucial en estudios de investigación y análisis estadístico.

La comparación de series de datos binarios no solo es relevante desde un punto de vista teórico, sino que también tiene importantes implicaciones prácticas. En campos como la medicina, la psicología y el análisis de mercado, la capacidad de comparar varianzas y proporciones puede ayudar a los académicos a tomar decisiones informadas basadas en datos cuantitativos. Al identificar patrones y diferencias significativas, se pueden desarrollar estrategias más efectivas y precisas para abordar problemas específicos.

Es fundamental considerar las líneas futuras de investigación en este ámbito. La evolución de técnicas estadísticas y el aumento del acceso a grandes volúmenes de datos binarios abren nuevas oportunidades para profundizar en la comparación de series. El desarrollo de métodos más sofisticados y la aplicación de modelos estadísticos avanzados permitirán una comprensión más profunda de la variabilidad y las relaciones entre variables

en contextos binarios. Además, la integración de herramientas de aprendizaje automático podría ofrecer nuevas perspectivas en la identificación de patrones complejos que aún no han sido completamente explorados.

3.2 Comparativa de Pruebas Estadísticas en Machine Learning: McNemar y Q de Cochran

El campo del machine learning ha experimentado un crecimiento exponencial en los últimos años, transformándose en una herramienta esencial en diversas industrias. Aquí es donde la estadística juega un papel fundamental, proporcionando las herramientas necesarias para interpretar resultados, medir la efectividad de los modelos y garantizar que las conclusiones extraídas sean válidas y fiables.

La estadística nos permite entender la variabilidad y la incertidumbre inherentes en los datos. En machine learning, los algoritmos se entrenan utilizando conjuntos de datos, y es vital saber cómo evaluar su desempeño. Las pruebas estadísticas ayudan a los académicos y a los científicos de datos a determinar si las diferencias observadas entre distintos modelos o entre predicciones y resultados reales son significativas o simplemente el resultado de la aleatoriedad. En este sentido, el uso de pruebas estadísticas no solo mejora la robustez de los modelos, sino que también respalda la toma de decisiones basadas en datos.

La prueba de McNemar es una herramienta estadística que se utiliza para evaluar la diferencia en proporciones de dos grupos relacionados. Es especialmente útil en situaciones donde se desea comparar dos clasificaciones o predicciones de un mismo conjunto de datos, como en el caso de modelos de clasificación que se evalúan en las mismas observaciones (Roldán et al., 2024). Esta prueba es ideal para datos binarios y proporciona una forma de

identificar si hay una diferencia significativa en el rendimiento de dos modelos.

Por otro lado, la Q de Cochran es una extensión de la prueba de McNemar, diseñada para evaluar la homogeneidad de proporciones en más de dos grupos relacionados. Es particularmente útil cuando se analizan datos categóricos en experimentos donde se desea comparar múltiples tratamientos o condiciones (Rayner y Livingston, 2022). Esta prueba permite a los académicos determinar si hay diferencias significativas en las proporciones entre varios modelos o clasificaciones, facilitando así una evaluación más completa de los resultados. La comprensión de estas pruebas estadísticas es esencial para cualquier profesional que trabaje en machine learning, ya que permite validar modelos y tomar decisiones informadas basadas en un análisis riguroso de los datos.

La prueba de McNemar es una prueba estadística no paramétrica utilizada para analizar datos de tablas de contingencia 2x2, especialmente en situaciones donde se evalúan cambios en respuestas categóricas de los mismos sujetos en dos momentos diferentes o bajo dos condiciones distintas. Esta prueba es particularmente útil en estudios de diseño pareado, donde cada sujeto actúa como su propio control. Para ilustrar, al comparar el rendimiento de dos modelos de machine learning en una misma base de datos, la prueba de McNemar permite evaluar si hay diferencias significativas en el número de clasificaciones correctas entre ambos modelos.

La hipótesis nula de la prueba establece que no hay diferencia en las proporciones de respuestas entre los dos grupos comparados, aun cuando la hipótesis alternativa sugiere que sí existe una diferencia. La prueba se centra en las discordancias de las respuestas, es decir, las instancias en las que un

modelo clasifica incorrectamente en un caso y el otro modelo clasifica correctamente.

En el contexto de machine learning, la prueba de McNemar se emplea comúnmente para evaluar el desempeño comparativo de diferentes algoritmos o modelos. Esto es, supongamos que se entrenan dos modelos para un problema de clasificación binaria, como la detección de spam en correos electrónicos. Tras realizar las predicciones sobre un conjunto de datos de prueba, se puede aplicar la prueba de McNemar para determinar si hay una diferencia estadísticamente significativa en el número de correos clasificados correcta o incorrectamente por cada modelo. Esta prueba es especialmente valiosa cuando se dispone de un conjunto de datos limitado y se busca maximizar la información obtenida a partir de él. Además, es útil en escenarios de ajuste y validación de modelos, donde se quiere decidir si un modelo nuevo ofrece mejoras significativas sobre uno previamente establecido.

Los resultados de la prueba de McNemar se expresan a través de un valor p, que indica la probabilidad de observar los resultados obtenidos, o más extremos, bajo la hipótesis nula. Un valor p inferior a un nivel de significancia predefinido (generalmente 0.05) sugiere que se debe rechazar la hipótesis nula, indicando que hay evidencia suficiente para afirmar que hay diferencias significativas en el rendimiento de los modelos comparados.

Además, es importante tener en cuenta el tamaño del efecto, que proporciona información adicional sobre cuán relevante es la diferencia observada. Aun cuando se obtenga un valor p significativo, un tamaño de efecto pequeño podría indicar que la diferencia en el rendimiento no es de gran importancia práctica. Por lo tanto, la prueba de McNemar no solo ayuda

a determinar si hay diferencias, sino que también permite contextualizar esos hallazgos dentro de la aplicación práctica del machine learning.

La Q de Cochran es una prueba estadística no paramétrica que se utiliza para evaluar la homogeneidad de proporciones en estudios donde se tienen múltiples grupos o tratamientos. Es especialmente útil en situaciones donde se desea comparar resultados categóricos a través de varias muestras relacionadas. Esta prueba se aplica a datos en formato de tablas de contingencia 2xk, donde "k" representa el número de grupos o tratamientos. En el contexto de machine learning, la Q de Cochran se convierte en una herramienta valiosa cuando se analizan modelos que generan predicciones categóricas a partir de diferentes conjuntos de datos o condiciones experimentales (Rayner y Livingston, 2022).

A diferencia de la prueba de McNemar, que se centra en comparar dos proporciones de un mismo grupo de sujetos antes y después de un tratamiento o en dos condiciones diferentes, la Q de Cochran permite la comparación de más de dos grupos. Entre tanto que la prueba de McNemar se utiliza principalmente para evaluar cambios en clasificaciones binarias, la Q de Cochran se expande a situaciones donde hay múltiples condiciones o tratamientos, facilitando el análisis comparativo de modelos de machine learning que están siendo evaluados en diferentes escenarios.

La Q de Cochran se puede aplicar en varios escenarios dentro del ámbito del machine learning. Por ejemplo, cuando se quiere evaluar la eficacia de diferentes algoritmos de clasificación en un mismo conjunto de datos, se puede utilizar esta prueba para determinar si hay diferencias significativas en las tasas de clasificación correcta entre los modelos. Asimismo, en estudios de validación cruzada donde se comparan múltiples configuraciones de

hiperparámetros, la Q de Cochran puede ayudar a establecer cuál de las configuraciones ofrece un rendimiento superior en términos de precisión o recuperación.

Además, en aplicaciones de ensemble learning, donde se combinan las predicciones de varios modelos, la Q de Cochran puede ser útil para analizar si las diferencias en las predicciones de los modelos individuales son estadísticamente significativas al combinar sus resultados. Esto permite a los académicos y practicantes de machine learning tomar decisiones informadas sobre la selección de modelos y la configuración de sus experimentos, asegurando que las conclusiones sean respaldadas por un análisis estadístico sólido.

La prueba de McNemar es particularmente útil en situaciones donde se desea comparar dos clasificaciones relacionadas, es decir, cuando se tiene un mismo conjunto de individuos o muestras que se han evaluado en dos ocasiones diferentes (Roldán et al., 2024). Pongamos el caso de, en el contexto del machine learning, puede ser aplicable al evaluar el rendimiento de dos modelos diferentes en un conjunto de datos específico. Esta prueba se utiliza comúnmente en tareas de clasificación binaria, donde se pueden observar las decisiones correctas e incorrectas de los modelos. Se recomienda su uso cuando se cuenta con un diseño de estudio en el que los pares de observaciones son dependientes, como en estudios antes y después, o en comparaciones de dos modelos que se aplican a las mismas instancias.

Por otro lado, la Q de Cochran es más adecuada para situaciones en las que se evalúan múltiples tratamientos o condiciones sobre el mismo grupo de sujetos. En el ámbito del machine learning, esta prueba es ideal cuando se comparan más de dos modelos de clasificación en función de un mismo

conjunto de datos. La Q de Cochran puede ser utilizada para examinar si existe una diferencia significativa en la tasa de aciertos entre varios algoritmos, permitiendo así un análisis más completo en estudios donde se implementan múltiples clasificaciones. Esta prueba es especialmente relevante en entornos donde se necesitan comparar varios modelos que se analizan bajo las mismas condiciones, garantizando que las decisiones se toman sobre la misma base de datos.

La elección entre la prueba de McNemar y la Q de Cochran depende de varios factores. En primer lugar, es fundamental considerar el número de modelos o tratamientos que se están comparando. Si solo se están comparando dos modelos, la prueba de McNemar es la opción más adecuada. Sin embargo, si se están evaluando tres o más modelos, la Q de Cochran se convierte en la herramienta preferida. Otro aspecto a tener en cuenta es la naturaleza de los datos; ambos métodos requieren que las observaciones sean dependientes, pero la Q de Cochran puede ofrecer una mayor flexibilidad al manejar múltiples categorías (Rayner y Livingston, 2022).

Además, es importante considerar la distribución de los datos y la naturaleza de los errores de clasificación. Entre tanto que la prueba de McNemar se basa en una tabla de contingencia 2x2 que examina los cambios en las clasificaciones, la Q de Cochran requiere una extensión a múltiples dimensiones, lo que puede complicar el análisis si los datos no se distribuyen adecuadamente. Por último, la interpretación de los resultados también debe ser un factor en la elección: la prueba de McNemar proporciona un enfoque más directo y simple, mientras que la Q de Cochran puede ofrecer una visión más amplia y compleja sobre las relaciones entre múltiples variables.

La selección entre la prueba de McNemar y la Q de Cochran debe basarse en las características específicas del estudio, incluyendo el número de modelos a comparar, la estructura de los datos y los objetivos del análisis. En el ámbito del machine learning, la estadística juega un papel fundamental en la evaluación y validación de modelos. La prueba de McNemar y la Q de Cochran son dos herramientas estadísticas que permiten analizar la efectividad de modelos de clasificación en situaciones donde se dispone de datos emparejados o múltiples clasificaciones de un conjunto de muestras.

La prueba de McNemar se destaca por su simplicidad y su capacidad para comparar la efectividad de dos modelos en clasificaciones binarias, siendo especialmente útil en contextos donde se requiere evaluar cambios en el rendimiento entre dos enfoques diferentes. Por otro lado, la Q de Cochran se presenta como una opción robusta para escenarios en los que se necesita comparar más de dos modelos o condiciones, permitiendo una visión más amplia sobre la consistencia de las clasificaciones en un conjunto de datos.

La elección entre estas dos pruebas debe basarse en el contexto específico del estudio, considerando factores como el número de modelos a comparar, la naturaleza de los datos y el tipo de resultados que se desean obtener. Mientras que la prueba de McNemar es ideal para comparaciones directas, la Q de Cochran ofrece una solución más versátil en situaciones más complejas.

Tanto la prueba de McNemar como la Q de Cochran son herramientas valiosas en el arsenal de un científico de datos, su correcta aplicación puede proporcionar avances significativos sobre la efectividad de los modelos de machine learning, mejorando así la toma de decisiones basadas en datos y contribuyendo al avance del campo (Roldán et al., 2024). Al comprender cómo

y cuándo utilizar estas pruebas, los profesionales pueden maximizar el rendimiento de sus modelos y asegurar que sus conclusiones sean estadísticamente válidas y relevantes.

3.3 Comparativa de la Prueba de Fisher y la Prueba de Levene en Aprendizaje Automático: Aplicaciones, Limitaciones y Consideraciones Prácticas

El aprendizaje automático es un campo en constante evolución que combina la estadística, la informática y la teoría de la información para permitir que las máquinas aprendan de los datos y realicen predicciones o decisiones basadas en ellos. En este contexto, las pruebas estadísticas juegan un papel fundamental, ya que ayudan a los científicos de datos a validar sus modelos, evaluar su rendimiento y garantizar que las inferencias realizadas sean robustas y confiables.

Las pruebas estadísticas son herramientas que permiten determinar si las observaciones realizadas en un conjunto de datos son significativas o si podrían ser el resultado del azar. En el ámbito del aprendizaje automático, estas pruebas son cruciales para evaluar supuestos, como la normalidad de los datos o la homogeneidad de las varianzas, que son condiciones necesarias para el correcto funcionamiento de muchos algoritmos de modelado.

Ambas pruebas se utilizan para evaluar la homogeneidad de las varianzas, lo que es esencial para garantizar que los modelos sean precisos y generalizables. A través de este análisis, esperamos proporcionar a los lectores una comprensión más clara de cómo las pruebas estadísticas pueden mejorar la calidad del aprendizaje automático y contribuir a la toma de decisiones informadas en el análisis de datos.

La prueba de Fisher, también conocida como la prueba de la razón de verosimilitud de Fisher, es una técnica estadística utilizada para evaluar si hay una diferencia significativa entre las varianzas de dos o más grupos. Su fundamento se basa en la distribución F, que compara la variabilidad entre grupos con la variabilidad dentro de los grupos. Es especialmente útil en contextos donde se asume que los datos siguen una distribución normal y se busca determinar si las varianzas son homocedásticas, es decir, si son iguales entre los diferentes grupos.

El propósito principal de la prueba de Fisher es proporcionar a los académicos una herramienta para validar supuestos en análisis de varianza (ANOVA) y otros modelos estadísticos. Al establecer si las varianzas son iguales, los científicos de datos pueden tomar decisiones informadas sobre el uso de ciertos modelos y técnicas de aprendizaje automático que dependen de esta suposición.

En el ámbito del aprendizaje automático, la prueba de Fisher se utiliza comúnmente en la fase de preprocesamiento de datos. Así como, al elegir características para un modelo, es esencial comprender la variabilidad de las características en relación con las clases objetivo. Si se encuentran diferencias significativas en las varianzas de las características entre diferentes clases, esto puede indicar que ciertas características son más informativas y, por lo tanto, más relevantes para el modelo (Bolboacă et al., 2011).

Además, la prueba de Fisher se aplica en la evaluación de modelos, especialmente en el contexto de la validación cruzada. Al comparar el rendimiento de diferentes modelos, los científicos de datos pueden utilizar la prueba para determinar si las diferencias en la precisión o en otras métricas de rendimiento son estadísticamente significativas, lo que les permite

seleccionar el modelo más adecuado para la tarea en cuestión. A pesar de sus amplias aplicaciones, la prueba de Fisher tiene sus limitaciones. En primer lugar, su aplicabilidad está restringida a datos que cumplen con el supuesto de normalidad y homocedasticidad. Si los datos se desvían significativamente de estos supuestos, la prueba puede no ser confiable.

Otro punto a considerar es que la prueba de Fisher es sensible a los tamaños de muestra, pues, en situaciones donde hay tamaños de muestra desiguales entre grupos, la prueba puede producir resultados sesgados. Además, la prueba no proporciona información sobre la magnitud de las diferencias en las varianzas, lo que puede ser crucial en ciertas aplicaciones. Es importante mencionar que la prueba de Fisher no debe utilizarse como una única herramienta de evaluación. Debe complementarse con otras pruebas y métodos estadísticos para obtener una comprensión más completa de los datos y sus características. Esto es especialmente relevante en el aprendizaje automático, donde la complejidad de los datos requiere un enfoque multifacético para una evaluación adecuada.

La prueba de Levene es una técnica estadística que se utiliza para evaluar la homogeneidad de las varianzas en diferentes grupos. A diferencia de la prueba de Bartlett, que asume que los datos provienen de distribuciones normales, la prueba de Levene es más robusta y no requiere esta suposición, lo que la convierte en una opción preferida en muchos escenarios prácticos. Su propósito principal es determinar si las varianzas de dos o más grupos son estadísticamente equivalentes. Esto es crucial en el ámbito del aprendizaje automático, ya que muchas técnicas de modelado, como la regresión y los análisis de varianza, asumen que las varianzas son homogéneas entre los grupos analizados.

En el contexto del aprendizaje automático, la prueba de Levene se puede aplicar en diversas etapas del proceso de modelado. En concreto, al comparar el rendimiento de diferentes modelos en función de un conjunto de métricas, es vital asegurarse de que las varianzas de estas métricas sean homogéneas. Si se observa que las varianzas son significativamente diferentes, esto puede indicar que uno o más modelos están funcionando de manera inconsistente, lo que podría afectar la interpretación de los resultados.

Además, la prueba de Levene puede ser útil en la validación cruzada, donde se evalúa la estabilidad de un modelo a través de múltiples particiones del conjunto de datos. Si las varianzas de las métricas de rendimiento son homogéneas, se puede tener mayor confianza en que el modelo generaliza bien. Por lo tanto, la prueba de Levene se convierte en una herramienta valiosa para los científicos de datos que buscan asegurar la validez de sus modelos. Existen varias pruebas que se utilizan para evaluar la homogeneidad de varianzas, incluyendo la prueba de Bartlett y la prueba de Brown-Forsythe. La principal diferencia entre la prueba de Levene y la prueba de Bartlett es que esta última es sensible a la normalidad de los datos. En situaciones donde los datos no siguen una distribución normal, la prueba de Levene es preferible debido a su robustez.

La prueba de Brown-Forsythe, por otro lado, es una modificación de la prueba de Levene que utiliza la mediana en lugar de la media para calcular las desviaciones. Esta modificación proporciona una mayor robustez ante la presencia de valores atípicos. En síntesis, a la vez que la prueba de Levene es una opción sólida y versátil, los científicos de datos deben considerar las características específicas de sus datos y el contexto del análisis al seleccionar la prueba más adecuada para evaluar la homogeneidad de varianzas.

La prueba de Fisher y la prueba de Levene son herramientas estadísticas utilizadas para evaluar la homogeneidad de las varianzas en diferentes grupos, pero difieren en sus enfoques y supuestos. La prueba de Fisher, también conocida como la prueba F, se basa en la relación entre las varianzas de dos grupos y asume que los datos siguen una distribución normal. Esto significa que es sensible a las desviaciones de la normalidad, lo que puede afectar la validez de los resultados en situaciones donde los datos no cumplen esta suposición (Bolboacă et al., 2011).

Por otro lado, la prueba de Levene es más robusta frente a violaciones de la normalidad y evalúa la igualdad de las varianzas utilizando las desviaciones absolutas de las medias o medianas de los grupos, lo que la hace menos susceptible a la influencia de valores atípicos. Esta característica la convierte en una opción preferida en muchos contextos de aprendizaje automático, donde los datos pueden no seguir distribuciones normales y pueden contener variaciones.

La elección entre la prueba de Fisher y la prueba de Levene depende del contexto específico del análisis y de las características de los datos. Si se tiene la certeza de que los datos son normales y se espera que no contengan variaciones significativas, la prueba de Fisher puede ser adecuada y ofrecer resultados precisos. Sin embargo, en la práctica del aprendizaje automático, donde la normalidad raramente se puede garantizar, la prueba de Levene es generalmente más recomendable.

Los científicos de datos deben considerar el tipo de modelo que están utilizando y la naturaleza de los datos al decidir qué prueba aplicar. Se puede citar, si se están evaluando modelos de regresión que asumen homogeneidad de varianzas, la prueba de Levene puede proporcionar una mejor evaluación

de los supuestos del modelo. En contraste, en situaciones donde se tiene un diseño experimental controlado con grupos bien definidos y se puede asumir normalidad, la prueba de Fisher podría ser considerada (Bolboacă et al., 2011).

La elección de la prueba estadística adecuada tiene implicaciones significativas en la interpretación de los resultados y en la toma de decisiones en el ámbito del aprendizaje automático. Utilizar la prueba de Fisher en datos que no cumplen con sus supuestos puede llevar a conclusiones erróneas y a un mal desempeño del modelo, a la vez que la prueba de Levene puede ofrecer un enfoque más confiable en escenarios diversos.

Además, es fundamental que los científicos de datos estén al tanto de las limitaciones de cada prueba y de la importancia de realizar un análisis previo de los datos. La validación de los supuestos de normalidad y homogeneidad de varianzas debe ser parte integral del proceso de modelado, y la elección de la prueba adecuada puede mejorar significativamente la robustez y la fiabilidad de los resultados en los proyectos de aprendizaje automático.

En el ámbito del aprendizaje automático, la correcta aplicación de pruebas estadísticas es fundamental para garantizar la validez y la fiabilidad de los modelos construidos. Tanto la prueba de Fisher como la prueba de Levene ofrecen herramientas valiosas para evaluar la homogeneidad de varianzas entre grupos, una consideración crucial en la mayoría de los análisis estadísticos. La prueba de Fisher, con su enfoque en la comparación de varianzas, es particularmente útil en situaciones donde se asume que los datos siguen una distribución normal. Su aplicación puede ser beneficiosa en modelos donde la normalidad de los datos es una premisa aceptada. Sin embargo, su sensibilidad a las desviaciones de esta suposición puede llevar a

conclusiones erróneas en contextos donde los datos no cumplen con la normalidad, lo que limita su uso en una variedad de escenarios prácticos.

Por otro lado, la prueba de Levene se destaca por su robustez frente a las violaciones de la normalidad, su capacidad para ofrecer pruebas más fiables en situaciones no normales la convierte en una herramienta esencial para los científicos de datos que trabajan con conjuntos de datos diversos y complejos (Neuhäuser y Schmitt, 2025). A. Además, su relevancia en la evaluación de modelos permite a los analistas tener una visión más clara sobre la homogeneidad de las varianzas, lo que puede influir significativamente en la selección y ajuste de modelos.

La elección entre la prueba de Fisher y la prueba de Levene debe basarse en la naturaleza de los datos y las suposiciones que se pueden realizar sobre ellos. Entre tanto que la prueba de Fisher puede ser preferida en situaciones controladas donde se cumple la normalidad, la prueba de Levene ofrece una alternativa más flexible y robusta para una amplia gama de aplicaciones en aprendizaje automático. Al comprender las diferencias y aplicaciones de cada prueba, los científicos de datos pueden tomar decisiones más informadas que mejoren la calidad y la efectividad de sus modelos.

Capítulo IV

Exploración de Pruebas de Asociación y Series de Tiempo: Fundamentos, Métodos y Aplicaciones

En el ámbito de la estadística y el análisis de datos, las pruebas de asociación y las series de tiempo son herramientas fundamentales que permiten a los académicos y analistas comprender mejor las relaciones entre variables y cómo estas evolucionan a lo largo del tiempo. Estas metodologías no solo son esenciales para el análisis descriptivo, sino que también desempeñan un papel crucial en la predicción y la toma de decisiones informadas en diversas disciplinas, incluyendo la economía, la biología, la sociología y la ingeniería.

Las pruebas de asociación se centran en identificar y evaluar la existencia de relaciones entre dos o más variables. Estas relaciones pueden ser directas o indirectas, y las pruebas estadísticas permiten determinar si estas asociaciones son significativas desde el punto de vista estadístico. Esto es, en la investigación social, un analista puede querer saber si existe una correlación entre el nivel educativo y el ingreso económico. A través de distintas pruebas de asociación, es posible cuantificar y validar estas relaciones, proporcionando así información valiosa para la formulación de políticas y estrategias.

Por otro lado, las series de tiempo se ocupan del análisis de datos recogidos a lo largo del tiempo, con el objetivo de identificar patrones, tendencias y ciclos. Este enfoque es especialmente útil en contextos donde las

variables están influenciadas por factores temporales, como en el caso de las ventas de un producto a lo largo de los meses o el cambio en las temperaturas a lo largo de las estaciones. El análisis de series de tiempo permite a los académicos no solo entender el comportamiento pasado de una variable, sino también hacer proyecciones futuras basadas en datos históricos.

4.1 Pruebas de asociación

Las pruebas de asociación son herramientas estadísticas fundamentales que nos permiten explorar y determinar la relación entre dos o más variables. Estas pruebas son esenciales en diversas disciplinas, como la psicología, la sociología, la biología y la economía, ya que ayudan a los académicos a identificar patrones, tendencias y correlaciones en los datos.

Las pruebas de asociación se utilizan para evaluar si existe una relación significativa entre variables categóricas o continuas. Su principal objetivo es determinar si las variaciones en una variable están relacionadas con las variaciones en otra. A saber, un investigador puede estar interesado en saber si existe una relación entre el nivel educativo y los ingresos de una población. Las pruebas de asociación permiten cuantificar la fuerza y dirección de esta relación, proporcionando así una base empírica para tomar decisiones y realizar inferencias. Existen diversas pruebas de asociación, cada una con características específicas adecuadas para diferentes tipos de datos y situaciones.

La prueba de Chi-cuadrado es una de las pruebas de asociación más utilizadas para analizar la relación entre dos variables categóricas, esta prueba evalúa si la distribución observada de una variable se desvía significativamente de la distribución esperada en función de la otra variable (Shimizu y Hashiguchi, 2024). Se calcula utilizando una tabla de contingencia,

donde se comparan las frecuencias observadas con las frecuencias esperadas. Si el valor del estadístico Chi-cuadrado es mayor que el valor crítico de la tabla, se puede concluir que existe una asociación significativa entre las variables.

La prueba de correlación de Pearson se utiliza para medir la fuerza y dirección de la relación lineal entre dos variables continuas. Esta prueba se basa en el coeficiente de correlación de Pearson (r), que varía entre -1 y 1. Un valor de r cercano a 1 indica una correlación positiva fuerte, entre tanto que un valor cercano a -1 indica una correlación negativa fuerte. Un valor de r cercano a 0 sugiere que no hay relación lineal entre las variables. Es importante tener en cuenta que la correlación no implica causalidad; es decir, si bien dos variables estén correlacionadas, no significa que una cause la otra.

La prueba de correlación de Spearman es una alternativa no paramétrica a la prueba de correlación de Pearson. Se utiliza cuando los datos no cumplen con los supuestos de normalidad o cuando las variables son ordinales. La prueba de Spearman evalúa la relación entre dos variables al convertir los datos en rangos y calcular el coeficiente de correlación de rangos de Spearman (ϱ). Al igual que con la prueba de Pearson, los valores de ϱ varían entre -1 y 1, proporcionando información sobre la dirección y fuerza de la asociación entre las variables (Lyerly, 1952).

Las pruebas de asociación tienen diversas aplicaciones en el análisis de datos. Así como, en el ámbito de la salud pública, se pueden utilizar para investigar la relación entre factores de riesgo y la prevalencia de enfermedades. En el campo de la educación, se pueden analizar las correlaciones entre el rendimiento académico y las variables socioeconómicas. Asimismo, en el marketing, estas pruebas ayudan a identificar las preferencias

de los consumidores en función de diferentes características demográficas. Las pruebas de asociación son herramientas esenciales para comprender las relaciones entre variables y son ampliamente aplicadas en diversas áreas del conocimiento. Con su uso adecuado, los académicos pueden extraer conclusiones significativas y contribuir al avance del conocimiento en sus respectivos campos.

Las series de tiempo son conjuntos de datos que se recogen en secuencias temporales, donde las observaciones están ordenadas cronológicamente. Este tipo de datos es fundamental en diversas disciplinas, como la economía, la meteorología, la ingeniería y la investigación social, ya que permite analizar y predecir comportamientos futuros a partir de patrones históricos. Una serie de tiempo se define como una colección de observaciones de una variable en intervalos de tiempo sucesivos. Las características principales de las series de tiempo incluyen:

i. *Tendencia:* Se refiere a la dirección general que sigue la serie a lo largo del tiempo, que puede ser ascendente, descendente o constante.
ii. *Estacionalidad:* Son patrones que se repiten en intervalos regulares, como cambios estacionales en las ventas de un producto.
iii. *Ciclos:* Fluctuaciones a largo plazo que no son regulares y que pueden estar relacionadas con factores económicos, sociales o políticos.
iv. *Ruido:* Variaciones aleatorias e impredecibles que pueden interferir con el comportamiento de la serie.

La identificación y análisis de estas características son fundamentales para la modelización de series de tiempo:

i. *Modelo ARIMA (AutoRegressive Integrated Moving Average):* Este modelo combina componentes autoregresivos y de media móvil, junto con la diferenciación de la serie para hacerla estacionaria. El ARIMA es ampliamente utilizado debido a su flexibilidad y capacidad para modelar una variedad de patrones de series de tiempo, incluyendo tendencias y estacionalidades (Pereira et al., 2024).

ii. *Modelo de suavizamiento exponencial:* Este enfoque se basa en la idea de que las observaciones más recientes deberían tener un mayor peso en la previsión futura. Los modelos de suavizamiento exponencial son especialmente útiles para series de tiempo con patrones estacionales y se utilizan frecuentemente en pronósticos de corto plazo.

iii. *Modelo de regresión con errores auto correlacionados:* Este modelo combina la regresión lineal con la consideración de que los errores en la predicción están correlacionados en el tiempo. Es útil en situaciones donde se sospecha que la variabilidad de los errores está relacionada con el tiempo y permite obtener previsiones más precisas.

Las series de tiempo tienen aplicaciones en múltiples áreas. Algunas de las más relevantes incluyen:

- Economía: Análisis de indicadores económicos como el PIB, tasas de desempleo, y precios de acciones a lo largo del tiempo para evaluar el rendimiento económico y realizar predicciones.

- Meteorología: Predicción del clima mediante el análisis de datos históricos de temperatura, precipitaciones y otros fenómenos atmosféricos.

- Salud pública: Monitoreo de enfermedades y brotes epidémicos al analizar la incidencia de casos a lo largo del tiempo para identificar patrones y prever futuros brotes.

- Marketing: Análisis de ventas de productos a lo largo del tiempo, identificando tendencias y patrones estacionales que pueden informar decisiones de inventario y estrategias de marketing.

Las series de tiempo son herramientas valiosas para el análisis y la predicción en diversas disciplinas, permitiendo a los académicos y profesionales identificar patrones y hacer pronósticos informados basados en datos históricos. Tanto las pruebas de asociación como las series de tiempo son herramientas estadísticamente relevantes que desempeñan un papel crucial en el análisis de datos en diversas disciplinas. Las pruebas de asociación permiten a los académicos identificar relaciones y patrones significativos entre variables, proporcionando una base sólida para la inferencia estadística. Herramientas como la prueba de Chi-cuadrado, las correlaciones de Pearson y Spearman son fundamentales para explorar y validar estas relaciones, lo que es especialmente útil en campos como la psicología, la medicina y las ciencias sociales (Lyerly, 1952).

Por otro lado, el análisis de series de tiempo se enfoca en la evolución y comportamiento de datos a lo largo del tiempo. Este enfoque es esencial para la predicción y el modelado de tendencias en áreas como la economía, la meteorología y la planificación de negocios. Modelos como ARIMA, el suavizamiento exponencial y la regresión con errores auto correlacionados ofrecen diferentes perspectivas y metodologías para abordar problemas complejos relacionados con la temporalidad de los datos.

Ambos enfoques, si bien distintos en su aplicación y metodología, comparten el objetivo común de proporcionar información valiosa y basada en evidencia para la toma de decisiones informadas. La integración de pruebas de asociación y análisis de series de tiempo puede enriquecer aún más la comprensión de los fenómenos estudiados, permitiendo a los académicos y profesionales obtener una visión más completa y matizada de los datos que analizan.

4.2 Análisis de Homogeneidad en Series Temporales y Aplicación del Modelo Multiplicativo Estacional de Holt-Winters

El análisis de series temporales se ha convertido en una herramienta fundamental en diversos campos, como la economía, la meteorología y la salud pública, entre otros. En este contexto, la homogeneidad de las series temporales juega un papel esencial, ya que asegura que los datos sean consistentes y comparables a lo largo del tiempo. La falta de homogeneidad puede llevar a interpretaciones erróneas y resultados engañosos. Por otro lado, el modelo multiplicativo estacional de Holt-Winters se presenta como una de las metodologías más efectivas para modelar y pronosticar series temporales con patrones estacionales. Este modelo no solo captura la tendencia y la estacionalidad, sino que también se adapta a variaciones en la amplitud de las fluctuaciones, ofreciendo una solución robusta para los analistas de datos (Alonso et al., 2021).

La homogeneidad en series temporales es un concepto fundamental que se refiere a la consistencia y estabilidad de las propiedades estadísticas de una serie a lo largo del tiempo. La falta de homogeneidad puede derivar de

cambios en la estructura de la serie, lo que influye significativamente en el análisis y la interpretación de los datos.

La homogeneidad en series temporales implica que las características estadísticas de la serie, como la media, la varianza y la covarianza, permanecen constantes a lo largo del tiempo. En otras palabras, una serie temporal es homogénea si sus propiedades no cambian en diferentes intervalos de tiempo. Este concepto es crucial porque muchas técnicas de análisis y modelado, como la regresión y el análisis de series temporales, suponen que los datos son homogéneos. Si una serie presenta heterogeneidad, se puede requerir un enfoque diferente para su análisis, ya que los resultados pueden ser engañosos o incorrectos. Existen diversos métodos para evaluar la homogeneidad de una serie temporal. Entre los más utilizados se encuentran:

i. *Pruebas de cambio de media:* Estas pruebas, como la prueba de Chow, evalúan si hay diferencias significativas en la media de la serie en diferentes períodos. Un cambio significativo puede indicar una falta de homogeneidad.
ii. *Pruebas de varianza:* Métodos como la prueba de Levene o la prueba de Bartlett se utilizan para determinar si la varianza de los datos es constante a lo largo del tiempo. La variación en la varianza puede señalar heterogeneidad en la serie.
iii. *Análisis gráfico:* Los gráficos de series temporales, como los gráficos de control y los diagramas de dispersión, permiten visualizar cambios en la tendencia, la estacionalidad y la variabilidad, ayudando a identificar posibles periodos de heterogeneidad.
iv. *Modelos de cambio estructural:* Estos modelos, como los modelos ARIMA con cambios estructurales, permiten identificar y modelar

puntos de cambio en la serie, proporcionando una forma de abordar la heterogeneidad.

La homogeneidad es esencial en el análisis de series temporales por varias razones. Primero, permite la validez de los supuestos de muchos métodos estadísticos. Si una serie es homogénea, se pueden aplicar modelos lineales y realizar inferencias estadísticas confiables. En segundo lugar, la identificación de períodos no homogéneos puede ayudar a los analistas a entender mejor los factores que afectan la serie, facilitando la toma de decisiones informadas. Por último, la falta de homogeneidad puede llevar a conclusiones erróneas y a una mala predicción del comportamiento futuro de la serie, lo que puede tener implicaciones significativas en campos como la economía, la meteorología y la planificación empresarial.

El modelo multiplicativo estacional de Holt-Winters es una técnica ampliamente utilizada en el análisis de series temporales, especialmente para datos que presentan patrones estacionales. Este modelo se basa en la descomposición de la serie temporal en tres componentes principales: nivel, tendencia y estacionalidad. Al combinar estas tres componentes mediante una estructura multiplicativa, el modelo logra capturar la complejidad de los datos estacionales de forma efectiva. El modelo multiplicativo de Holt-Winters se puede describir mediante las siguientes ecuaciones:

i. *Nivel (L):* Representa el valor promedio de la serie en un momento dado.
ii. *Tendencia (T):* Refleja el cambio en el nivel a lo largo del tiempo, indicando si hay un aumento o disminución en la serie.
iii. *Estacionalidad (S):* Captura los patrones estacionales que se repiten en intervalos regulares, multiplicando el nivel y la tendencia.

Las ecuaciones del modelo se definen como sigue:

- $L_t = \alpha \frac{Y_t}{S_{t-m}} + (1 - \alpha)(L_{t-1} + T_{t-1})$

- $T_t = \beta (L_t - L_{t-1}) + (1 - \beta)T_{t-1}$

- $S_t = \gamma \frac{Y_t}{L_t} + (1 - \gamma)S_{t-m}$

Donde Y_t es el valor observado en el tiempo t, α, β y γ son los parámetros de suavizamiento para el nivel, la tendencia y la estacionalidad, respectivamente, y m es el número de períodos en una temporada. El modelo multiplicativo de Holt-Winters se aplica comúnmente en diversas áreas, como la economía, la meteorología y la planificación de la producción, donde la estacionalidad es un factor significativo. Para ilustrar, en el análisis de las ventas de un producto, el modelo puede ayudar a predecir el comportamiento futuro basándose en patrones de ventas pasadas que se repiten anualmente.

Para implementar el modelo, los analistas deben inicializar los componentes del nivel, la tendencia y la estacionalidad, y luego ajustar los parámetros de suavizamiento usando métodos de optimización, como el método de mínimos cuadrados. Una vez calibrado, el modelo puede generar pronósticos para períodos futuros, lo que permite a las empresas anticipar cambios en la demanda y ajustar sus estrategias en consecuencia.

El modelo multiplicativo estacional de Holt-Winters presenta varias ventajas. Entre ellas, se destaca su capacidad para manejar datos con variaciones estacionales significativas, lo que lo convierte en una herramienta potente para la previsión en contextos donde la estacionalidad es un factor determinante. Además, su estructura multiplicativa le permite modelar de

manera efectiva situaciones en las que la magnitud de la estacionalidad cambia con el nivel de la serie.

Sin embargo, el modelo también tiene sus desventajas. Por un lado, puede ser sensible a la elección de los parámetros de suavizamiento, lo que puede influir en la calidad del pronóstico. Además, el modelo asume que las relaciones entre las componentes son constantes a lo largo del tiempo, lo que puede no ser cierto en series temporales más complejas o en presencia de cambios estructurales. Por último, el modelo puede no capturar adecuadamente las irregularidades o eventos atípicos que pueden impactar los datos. El modelo multiplicativo estacional de Holt-Winters es una herramienta valiosa para el análisis de series temporales estacionales, pero su aplicación debe hacerse con cuidado y consideración de las características específicas de los datos en cuestión.

La interpretación de los resultados en el contexto de la prueba de homogeneidad y la aplicación del modelo multiplicativo estacional de Holt-Winters es fundamental para comprender las dinámicas de las series temporales analizadas. Al realizar la prueba de homogeneidad, se busca determinar si la serie presenta cambios estructurales que puedan afectar su comportamiento a lo largo del tiempo. Los resultados de estas pruebas, que pueden incluir estadísticas de test y gráficos de series, permiten identificar segmentos homogéneos y, por ende, establecer intervalos de tiempo en los que la serie puede ser tratada como consistente.

Por otro lado, la implementación del modelo de Holt-Winters en series temporales con estacionalidad es crucial para la proyección futura de datos. Los resultados obtenidos de este modelo, que incluyen pronósticos y medidas de ajuste como el error cuadrático medio (ECM), ofrecen una visión clara

sobre la capacidad predictiva del modelo. Cuando el modelo se ajusta adecuadamente a los datos, se puede observar una reducción en la variabilidad residual y un incremento en la precisión de las proyecciones, lo que valida la elección del enfoque multiplicativo.

A partir de la interpretación de los resultados, se evidencian áreas de mejora y nuevas direcciones para futuras investigaciones. En primer lugar, se sugiere profundizar en la exploración de diferentes métodos estadísticos para la prueba de homogeneidad, considerando técnicas más avanzadas que puedan captar cambios sutiles en la estructura de la serie. La incorporación de algoritmos de machine learning podría proporcionar una perspectiva innovadora y mejorar la robustez de los análisis.

Asimismo, se recomienda que futuros estudios evalúen la aplicabilidad del modelo de Holt-Winters en series temporales no solo estacionales, sino también en aquellas con tendencias o ciclos más complejos. Exploraciones sobre la combinación de modelos, como la integración de Holt-Winters con técnicas de suavizamiento o modelos ARIMA, podrían resultar en un aumento en la precisión de las predicciones.

La prueba de homogeneidad de series temporales y el modelo multiplicativo estacional de Holt-Winters juegan un papel crucial en el análisis y pronóstico de datos temporales. La homogeneidad es un requisito esencial para garantizar la validez de los modelos aplicados, aun cuando el modelo de Holt-Winters proporciona una metodología efectiva para tratar la estacionalidad (Alonso et al., 2021). Al integrar ambos enfoques, se obtiene una herramienta poderosa para la toma de decisiones informadas en diversos campos, desde la economía hasta la meteorología.

Sin embargo, es fundamental seguir investigando y adaptando las metodologías existentes para enfrentar la creciente complejidad de las series temporales modernas. La continua evolución de las técnicas estadísticas y computacionales promete enriquecer el análisis de datos en el futuro, contribuyendo así a un entendimiento más profundo de los fenómenos que se manifiestan en el tiempo.

4.3 metodología para Ajustar Modelos ARIMA en Series Temporales

Los modelos ARIMA, que significan "AutoRegressive Integrated Moving Average", son una de las herramientas más utilizadas en el análisis de series temporales. Estos modelos permiten capturar la estructura de dependencia temporal de los datos, facilitando así la predicción de valores futuros basándose en observaciones pasadas (Pereira et al., 2024).

Un modelo ARIMA se compone de tres elementos clave: la parte autorregresiva (AR), la parte de media móvil (MA) y la diferenciación (I). La parte autorregresiva se refiere a cómo los valores actuales de la serie dependen de sus valores pasados. La parte de media móvil modela el impacto de los errores pasados en el valor actual. La diferenciación se utiliza para hacer que la serie sea estacionaria, es decir, para eliminar tendencias y estacionalidades que pueden dificultar el análisis. La combinación de estos componentes permite a los analistas ajustar modelos que reflejan de manera precisa la dinámica de las series temporales.

La importancia de los modelos ARIMA radica en su capacidad para modelar y predecir series temporales que no son necesariamente estacionarias. En muchos contextos, los datos del mundo real presentan tendencias y patrones estacionales que complican el análisis. Los modelos

ARIMA ofrecen un marco robusto que puede adaptarse a estas características, permitiendo a los académicos y profesionales realizar pronósticos más precisos y fundamentados. Además, su flexibilidad y capacidad de ajuste los hacen aplicables en diversos campos, desde la economía hasta la meteorología, pasando por la ingeniería y la salud pública.

Los modelos ARIMA se utilizan ampliamente en diversas aplicaciones, como la previsión de ventas, el análisis financiero, la predicción de demanda energética y la modelización de datos climáticos. Por ejemplo, en el ámbito empresarial, una empresa puede utilizar un modelo ARIMA para prever las ventas de un producto en función de las tendencias pasadas y las fluctuaciones estacionales. En el sector financiero, se pueden aplicar para predecir precios de acciones o tasas de interés. La versatilidad de los modelos ARIMA los convierte en una herramienta invaluable para cualquier profesional que trabaje con datos temporales, permitiéndoles tomar decisiones más informadas basadas en análisis rigurosos.

Los modelos ARIMA son herramientas poderosas para el análisis de series temporales. Comprender los componentes que los conforman es fundamental para su correcta aplicación y ajuste. Un modelo ARIMA se compone de tres elementos principales: el componente autorregresivo (AR), el componente de media móvil (MA) y la diferenciación (I).

El componente autorregresivo del modelo ARIMA se refiere a la relación entre una observación y un número determinado de observaciones anteriores. En términos simples, se basa en la premisa de que los valores pasados de la serie temporal pueden utilizarse para predecir su valor futuro. Esto se expresa matemáticamente como una combinación lineal de los valores pasados, ponderados por ciertos coeficientes. El orden de la parte

autorregresiva, denotado como "p", indica cuántos valores pasados se toman en cuenta para realizar la predicción. En concreto, un modelo AR(1) solo considera el valor inmediatamente anterior, aun cuando que un modelo AR(2) toma en cuenta los dos valores anteriores.

El componente de media móvil, por otro lado, se basa en la relación entre una observación y los errores de predicción de observaciones anteriores. En este caso, el modelo ajusta la serie temporal considerando la influencia de las perturbaciones pasadas. El orden de la parte de media móvil, denotado como "q", indica cuántos errores pasados se utilizan en la predicción del valor actual. Al igual que en el caso del componente AR, un modelo MA(1) considera solo el error del periodo anterior, mientras que un MA(2) incluye los errores de los dos periodos anteriores.

La diferenciación es el proceso que permite transformar una serie temporal no estacionaria en una serie estacionaria. Una serie es estacionaria si sus propiedades estadísticas, como la media y la varianza, son constantes a lo largo del tiempo. Dado que muchos modelos de series temporales, incluido ARIMA, requieren estacionariedad, la diferenciación se utiliza para eliminar tendencias y estacionalidades. El orden de diferenciación se denota como "d" y representa el número de veces que se aplica la operación de diferenciación. Esto es, si se aplica una diferenciación simple, se restará el valor actual del valor anterior, creando una nueva serie que puede ser más adecuada para el modelado.

Los componentes autorregresivos, de media móvil y de diferenciación son esenciales para construir un modelo ARIMA sólido, cada uno de estos elementos juega un papel crucial en la captura de las dinámicas de la serie temporal, lo que permite realizar predicciones más precisas y efectivas

(Pereira et al., 2024). Conociendo y comprendiendo estos componentes, los analistas pueden ajustar modelos ARIMA que se adapten mejor a sus datos y necesidades específicas. El ajuste de un modelo ARIMA a una serie temporal implica una serie de pasos sistemáticos que permiten identificar, estimar y validar el modelo más adecuado para los datos en cuestión. Este proceso es fundamental para garantizar que las predicciones obtenidas sean precisas y confiables.

El primer paso en el proceso de ajuste de un modelo ARIMA es la identificación de la serie temporal. Esto implica examinar la serie para determinar sus características principales, como la tendencia, la estacionalidad y la estacionariedad.

i. *Análisis gráfico:* Se recomienda comenzar con gráficos como el gráfico de series temporales y el gráfico de autocorrelación (ACF) y autocorrelación parcial (PACF). Estos gráficos permiten visualizar el comportamiento de la serie y ayudan a detectar patrones y correlaciones.
ii. *Estacionariedad:* Un modelo ARIMA requiere que la serie temporal sea estacionaria, es decir, que sus propiedades estadísticas no cambien con el tiempo. Para evaluar la estacionariedad, se pueden utilizar pruebas estadísticas como la prueba de Dickey-Fuller aumentada (ADF) o la prueba KPSS. Si la serie no es estacionaria, es posible que se necesiten transformaciones, como la diferenciación, para alcanzar la estacionariedad.
iii. *Determinación de órdenes:* Una vez que la serie es estacionaria, se procede a identificar los órdenes del modelo ARIMA, que se representan como ARIMA(p, d, q), donde 'p' es el orden del componente autorregresivo, 'd' es el grado de diferenciación y 'q' es

el orden del componente de media móvil. Los gráficos ACF y PACF son útiles para determinar los valores de 'p' y 'q'.

Con los órdenes identificados, el siguiente paso es la estimación de los parámetros del modelo. Este proceso implica ajustar el modelo ARIMA a la serie temporal utilizando métodos estadísticos.

i. *Métodos de estimación:* Los métodos más comunes para la estimación de parámetros en modelos ARIMA son el método de mínimos cuadrados ordinarios y el método de máxima verosimilitud. Estas técnicas permiten encontrar los valores óptimos de los parámetros que minimizan el error en las predicciones.
ii. *Software estadístico:* Existen diversas herramientas y software estadísticos, como R, Python o EViews, que facilitan la implementación de modelos ARIMA y la estimación de sus parámetros. Estas plataformas ofrecen funciones específicas que automatizan el proceso y brindan resultados precisos.

Una vez que se ha ajustado el modelo, es crucial validar su adecuación antes de utilizarlo para realizar predicciones.

i. *Diagnóstico del modelo:* Se deben realizar pruebas de diagnóstico para evaluar la calidad del modelo ajustado. Esto incluye la verificación de los residuos del modelo, que deben comportarse como ruido blanco. Se pueden utilizar gráficos de residuos, así como pruebas como la prueba de Ljung-Box, para comprobar la independencia de los residuos.
ii. *Evaluación de la precisión:* Para evaluar la precisión del modelo, se pueden utilizar métricas como el error cuadrático medio (MSE), el error absoluto medio (MAE) o el error porcentual absoluto medio

(MAPE). Estas métricas permiten comparar el rendimiento del modelo ARIMA con otros modelos o enfoques de predicción.

iii. *Predicciones:* Es importante monitorear continuamente el rendimiento del modelo que se disponen de nuevos datos y hacer ajustes si es necesario, pues, el proceso de ajuste de un modelo ARIMA es fundamental para obtener pronósticos precisos y útiles. A través de la identificación, estimación y validación, se puede asegurar que el modelo se adapta adecuadamente a la serie temporal analizada, ofreciendo así una herramienta poderosa para el análisis de datos en el ámbito de las series temporales.

El uso de modelos ARIMA en el análisis de series temporales se ha consolidado como una herramienta fundamental para la predicción y el entendimiento de patrones en datos que varían a lo largo del tiempo. Su capacidad para capturar la dinámica subyacente de una serie mediante la combinación de componentes autorregresivos, de media móvil y diferenciación permite a los analistas obtener pronósticos más precisos y relevantes.

La versatilidad de los modelos ARIMA se traduce en su aplicabilidad en diversas disciplinas, desde la economía y la finanza hasta la meteorología y la ingeniería. Gracias a su robustez, los modelos ARIMA pueden adaptarse a diferentes tipos de series temporales, ya sean estacionarias o no, lo que amplía su utilidad en situaciones del mundo real donde los datos suelen presentar tendencias y patrones complejos.

Sin embargo, es crucial recordar que el éxito del ajuste de un modelo ARIMA depende en gran medida de una adecuada identificación de la serie temporal y de la correcta estimación de sus parámetros. Asimismo, la

validación del modelo es un paso esencial para garantizar que las predicciones sean fiables y útiles. La incorporación de técnicas de diagnóstico, como el análisis de los residuos, ayuda a mejorar la calidad de los modelos y a asegurar que se ajusten adecuadamente a los datos (Pereira et al., 2024). Los modelos ARIMA son una herramienta poderosa y esencial en el arsenal de los analistas de datos. Su capacidad para descomponer y entender series temporales complejas no solo facilita la toma de decisiones informadas, sino que también contribuye al avance de diversas áreas del conocimiento.

Conclusión

En el campo del aprendizaje automático, los métodos paramétricos juegan un papel decisivo, algoritmos como la regresión lineal y logística son ejemplos clásicos de modelos paramétricos que se utilizan para actividades de clasificación y predicción. Para ilustrar, un modelo de regresión logística puede ser empleado para predecir la probabilidad de que un estudiante apruebe un examen en función de variables como su asistencia, sus calificaciones previas y su participación en clase. Estos modelos son ampliamente utilizados debido a su simplicidad y a la facilidad con la que se interpretan los resultados, lo que permite a los educadores tomar decisiones informadas basadas en datos cuantitativos.

La integración de teorías estadísticas, tanto paramétricas como no paramétricas, en toda las áreas del saber representa un avance significativo en la forma en que se interpreta y se aborda el aprendizaje. En este libro, se ahondó sobre estas teorías, proporcionando herramientas esenciales para la evaluación de procesos varios y toma de decisiones informadas

Las teorías paramétricas, al ofrecer modelos basados en supuestos específicos sobre la distribución de los datos, permiten realizar análisis detallados y precisos que pueden ser fundamentales para medir el éxito, la intervención y la optimización de procesos. La aplicación en el aprendizaje automático ha mostrado resultados prometedores, evidenciando su relevancia en la creación de sistemas adaptativos que responden a las necesidades individuales y colectivas.

Por otro lado, las teorías no paramétricas emergen como una opción valiosa en contextos donde los supuestos sobre la distribución de los datos no se cumplen o son difíciles de establecer. Su flexibilidad y capacidad para manejar datos de diferentes formas demuestran cómo estos enfoques pueden descubrir patrones ocultos en el aprendizaje, lo que a su vez puede influir en la formulación de estrategias pedagógicas más efectivas. Los modelos no paramétricos son menos susceptibles a la influencia de valores atípicos, lo que puede ser crítico en cualquier área, donde algunas observaciones pueden ser extremas debido a factores contextuales o individuales, por lo que se obtienen estimaciones más robustas y representativas en los procesos a medir.

En conclusión, la elección entre estadística paramétrica y no paramétrica depende, en gran medida, de la naturaleza de los datos y los objetivos del análisis. Si los datos son continuos y cumplen con las condiciones de normalidad y homogeneidad de varianzas, las pruebas paramétricas son generalmente preferibles. Por el contrario, si los datos son ordinales, categóricos, o si hay dudas sobre la normalidad, las pruebas no paramétricas son más adecuadas. Además, es importante considerar el tamaño de la muestra. En muestras pequeñas, la estadística no paramétrica puede ser la mejor opción para evitar conclusiones erróneas. En contraste, con muestras grandes, donde el teorema central del límite sugiere que la distribución de la media se aproxima a la normalidad, las pruebas paramétricas pueden ser más ventajosas.

Bibliografía

Alexander, D.L., Tropsha, A., & Winkler, D.A. (2015). Beware of R(2): Simple, Unambiguous Assessment of the Prediction Accuracy of QSAR and QSPR Models. *Journal of chemical information and modeling*, *55*(7), 1316–1322. https://doi.org/10.1021/acs.jcim.5b00206

Alonso Brito, G.R., Rivero Villaverde, A., Lau Quan, A. *et al. (2021).* Comparison between SARIMA and Holt–Winters models for forecasting monthly streamflow in the western region of Cuba. *SN Appl. Sci.*, 3(671). https://doi.org/10.1007/s42452-021-04667-5

Bolboacă, S.D., Jäntschi, L., Sestraş, A.F., Sestraş, R.E., & Pamfil, D.C. (2011). Pearson-Fisher Chi-Square Statistic Revisited. *Information*, *2*(3), 528-545. https://doi.org/10.3390/info2030528

Burbano-Pantoja, V.M.A., Valdivieso-Miranda, M.A., y Burbano-Valdivieso, Á.S. (2022). Modelos estadísticos no paramétricos en los libros de texto del nivel universitario. *Revista de Investigación, Desarrollo e Innovación*, *12*(2), 265-278

Dagnino, J. (2014). Inferencia estadística: Pruebas de hipótesis. *Revista chilena de anestesia*, 2(43), 125-128. https://doi.org/10.25237/revchilanestv43n02.10

Flores-Ruiz, E., Miranda-Novales, M.G., y Villasís-Keever, M.A. (2017). El protocolo de investigación VI: cómo elegir la prueba estadística adecuada. Estadística inferencial. *Revista alergia México*, *64*(3), 364-370. https://doi.org/10.29262/ram.v64i3.304

Hart A. (2001). Mann-Whitney test is not just a test of medians: differences in spread can be important. *BMJ (Clinical research ed.)*, *323*(7309), 391–393. https://doi.org/10.1136/bmj.323.7309.391

Lanzagorta-Ortega, D., Carrillo-Pérez, D.L., y Carrillo-Esper, R. (2022). Inteligencia artificial en medicina: presente y futuro. *Gaceta médica de México*, *158*(Supl. 1), 17-21. https://doi.org/10.24875/gmm.m22000688

Lyerly, S.B. (1952) The average spearman rank correlation coefficient. *Psychometrika* 17, 421–428 (1952). https://doi.org/10.1007/BF02288917

Maestro-Gilmartín, L.M., Pozo-Pisabarro, R., García-Iglesias, M.J., Naveiro-Rilo, C., y Álvarez-Torices, J.C. (2017). Estudio descriptivo sobre la variación de la presión arterial en relación al trabajo a turnos en un servicio hospitalario de urgencias. *Revista de la Asociación Española de Especialistas en Medicina del Trabajo*, *26*(1), 69-75

Molina, M. (2025). Redes neuronales artificiales: fundamentos y aplicaciones. Evidencias en pediatría, 21(2), 1-6

Najem, R., Bahnasse, A., Fakhouri Amr, M. *et al.* (2025).Técnicas avanzadas de IA y big data en finanzas electrónicas: un estudio exhaustivo. *Discov Artif Intell*, 5, 102. https://doi.org/10.1007/s44163-025-00365-y

Neuhäuser, M., & Schmitt, S. (2025). A Note on the Robust Modification of the Ordered-Heterogeneity Test. *Stats*, *8*(2), 47. https://doi.org/10.3390/stats8020047

Paixão, P., Silva, N., Guerreiro, R.B., Blake, K., Bonelli, M., Morais, J.A.G., García-Arieta, A., & Gouveia, L.F. (2022). Evaluation of a Proposed Approach for the Determination of the Bioequivalence Acceptance Range for Narrow

Therapeutic Index Drugs in the European Union. *Pharmaceutics*, *14*(11), 2349. https://doi.org/10.3390/pharmaceutics14112349

Pereira da Veiga, C., Pereira da Veiga, C.R., Girotto, F.M. *et al.* (2024). Implementation of the ARIMA model for prediction of economic variables: evidence from the health sector in Brazil. *Humanit Soc Sci Commun,* 11(1068). https://doi.org/10.1057/s41599-024-03023-3

Rayner, J.C.W., & Livingston, G.C., Jr. (2022). Ordinal Cochran-Mantel-Haenszel Testing and Nonparametric Analysis of Variance: Competing Methodologies. *Stats*, *5*(4), 970-976. https://doi.org/10.3390/stats5040056

Rodríguez Flores, E.A., Garcés Giraldo, L., Valencia, J., y Valencia-Arias, A. (2025). Tendencias investigativas en el uso de técnicas de inteligencia artificial en la investigación científica. *Revista Venezolana De Gerencia*, *30*(109), 351-380. https://doi.org/10.52080/rvgluz.30.109.23

Roig-Vila, R. (2019). *Investigación e innovación en la Enseñanza Superior. Nuevos contextos, nuevas ideas.* Barcelona: Ediciones Octaedro

Roldán-Nofuentes, J.A., Sheth, T.S., & Vera-Vera, J.F. (2024). Hypothesis Test to Compare Two Paired Binomial Proportions: Assessment of 24 Methods. *Mathematics*, *12*(2), 190. https://doi.org/10.3390/math12020190

Shimizu, K., & Hashiguchi, H. (2024). Chi-Square Approximation for the Distribution of Individual Eigenvalues of a Singular Wishart Matrix. *Mathematics*, *12*(6), 921. https://doi.org/10.3390/math12060921

Taha, K. (2025). Análisis de Big Data en IoT, redes sociales, PLN y seguridad de la información: tendencias, desafíos y aplicaciones. *J Big Data,* 12(150). https://doi.org/10.1186/s40537-025-01192-9

Villegas Zamora, D.A. (2019). La importancia de la estadística aplicada para la toma de decisiones en Marketing. *Revista Investigación y Negocios*, *12*(20), 31-44

De esta edición de **“Estadística paramétrica y no paramétrica: Aplicaciones en inteligencia artificial y machine learning”**, se terminó de editar en la ciudad de Colonia del Sacramento en la República Oriental del Uruguay el 01 de agosto de 2025

EST. 2021 EMC
EDITORIAL MAR CARIBE

ESTADÍSTICA PARAMÉTRICA Y NO PARAMÉTRICA:

Aplicaciones en inteligencia artificial y machine learning

ISBN: 978-9915-698-24-3

www.editorialmarcaribe.es