MAESTRÍA EN CIENCIA S DE DATOS

Admisión

ASIGNATURAS CLAVE PARA LA FORMACIÓN PROFESIONAL EN LA MAESTRÍA

Ing. Wilson Gustavo Chango Ph.D.

2024

13/01/2024

CONTENIDO

Asignaturas clave para la formación profesional en la maestría. 3

1. **ANÁLISIS ESTADÍSTICO DE DATOS.** 3
2. **APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING)** 6
3. **ÉTICA Y PRIVACIDAD EN CIENCIA DE DATOS** 8
4. **BIG DATA Y COMPUTACIÓN EN LA NUBE** 10
5. **INTELIGENCIA ARTIFICIAL** 12

Bibliografía 15

Asignaturas clave para la formación profesional en la maestría.

1. Análisis estadístico de datos - Ofrece las bases fundamentales para comprender y analizar datos, esencial en cualquier disciplina relacionada con ciencia de datos.
2. Aprendizaje Automático (Machine Learning) - Proporciona habilidades avanzadas en técnicas de aprendizaje automático, que son cruciales para el procesamiento y análisis de grandes conjuntos de datos.
3. Ética y privacidad en ciencia de datos - Aborda los aspectos éticos y de privacidad que son esenciales en la gestión y uso responsable de la información.
4. Big Data y computación en la nube - Introduce conceptos y herramientas relacionadas con el manejo eficiente de grandes volúmenes de datos en entornos de computación en la nube.
5. Inteligencia artificial - Brinda una comprensión profunda de los principios y aplicaciones de la inteligencia artificial, siendo una competencia clave en el ámbito actual de la tecnología y la investigación.

Estas asignaturas se destacan debido a su relevancia en la formación profesional y técnica de los maestrantes en el programa. Sin embargo, es importante que la Comisión de Posgrado defina cuáles considera prioritarias según su criterio técnico específico.

1. **ANÁLISIS ESTADÍSTICO DE DATOS.**

- La mediana es **menos sensible a los valores atípicos** en comparación con la media. Si tienes un conjunto de datos con valores extremos, la media puede verse fuertemente influenciada, mientras que la mediana se mantiene más estable (Faraldo & Pateiro, 2013,p. 8).
- El principal objetivo del análisis estadístico descriptivo es **resumir, organizar y presentar los datos** de manera que sean comprensibles y útiles. Esto implica calcular medidas de tendencia central (como la media, la mediana y la moda),

medidas de dispersión (como la varianza y la desviación estándar), así como representar los datos mediante gráficos, tablas y otros métodos visuales (Orellana, 2001,p. 14)

- La varianza es un concepto estadístico fundamental que proporciona información acerca de la dispersión o **variabilidad de un conjunto de datos con respecto a su media**. En otras palabras, la varianza nos permite entender cuán dispersos están los datos alrededor de su valor central, que es la media. Cuando calculamos la varianza de un conjunto de datos, lo que estamos haciendo es determinar cuánto se alejan **individualmente cada uno de los puntos de datos de la media.** Si todos los puntos de datos están muy cerca de la media, la varianza será baja, indicando que hay poca dispersión. Por otro lado, si los puntos de datos se encuentran muy dispersos alrededor de la media, la varianza será alta, lo que significa que existe una mayor variabilidad entre los datos.
- Los **gráficos de barras son ideales para representar la distribución de frecuencias** de variables discretas, mientras que el histograma se utiliza más comúnmente para variables continuas. (Orellana, 2001,p.26)
- La mediana es una **medida de tendencia central adecuada para datos ordinales** porque se centra en el valor medio cuando los datos se ordenan (Aguilar Reyes et al., 2022,p. 48)
- La regresión lineal simple es una técnica estadística que se emplea con frecuencia en el análisis de datos para **entender y predecir la relación entre una variable dependiente y una variable independiente**. Aunque se denomina "simple" para indicar que involucra una sola variable independiente, es fundamental para establecer relaciones lineales básicas y comprender cómo una variable afecta a otra en términos de tendencia y dirección. El objetivo principal de la regresión

lineal simple es crear una línea recta que represente la relación más ajustada entre la variable dependiente y la variable independiente. Esta línea se determina utilizando técnicas matemáticas que minimizan la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por el modelo (Aguilar Reyes et al., 2022,p. 77)

- El concepto estadístico utilizado para determinar si la diferencia entre las medias de dos grupos es estadísticamente significativa se denomina **prueba de hipótesis** para dos muestras. Específicamente, una de las pruebas más comunes utilizadas con este propósito es la prueba t de Student para dos muestras independientes (Aguilar Reyes et al., 2022,p. 132)
- Un valor p menor a 0.05 en una prueba de hipótesis indica que los resultados observados son estadísticamente significativos a un nivel de significancia del 5%. En otras palabras, un valor $p < 0.05$ sugiere que hay evidencia suficiente para **rechazar la hipótesis nula (H0)** en favor de la hipótesis alternativa (Ha o H1) (Aguilar Reyes et al., 2022,p. 147).
- Un gráfico de torta es especialmente adecuado para **representar variables nominales**, que son categorías distintas sin un orden inherente. Las variables nominales se caracterizan por dividir los datos en diferentes categorías o grupos sin establecer un orden o jerarquía específica entre ellos. Al utilizar un gráfico de torta para variables nominales, puedes visualizar fácilmente cómo se distribuyen las diferentes categorías en relación con un total (Orellana, 2001,p.16)
- La correlación es una medida estadística que evalúa la relación lineal entre dos variables cuantitativas. **Esta medida proporciona información sobre la dirección y la fuerza de la relación entre las variables**, permitiendo determinar

cómo cambia una variable en respuesta a cambios en la otra variable (Aguilar Reyes et al., 2022,p. 73)

2. APRENDIZAJE AUTOMÁTICO (MACHINE LEARNING)

- Los árboles de decisión son una técnica de aprendizaje automático y un método de modelado predictivo ampliamente utilizados **para abordar problemas de clasificación de datos**. Estos árboles son representaciones gráficas de decisiones y se construyen a partir de datos de entrenamiento para segmentar o categorizar observaciones en diferentes clases o categorías basadas en características o variables predictoras (Pursky et al., 2018,p. 18).
- Cuando el modelo aprende directamente de los datos **sin intervención humana ni etiquetas predefinidas, se está utilizando un enfoque de aprendizaje automático denominado "aprendizaje no supervisado"**. En el aprendizaje no supervisado, los algoritmos procesan conjuntos de datos sin tener categorías o respuestas conocidas previamente para cada observación. El objetivo principal es descubrir patrones, estructuras, características o relaciones inherentes en los datos de forma autónoma, sin la guía de etiquetas o información externa proporcionada por humanos (Pursky et al., 2018,p. 26).
- Uno de los principales desafíos del sobreajuste en el aprendizaje automático es la tendencia de un modelo a ajustarse demasiado a los detalles y variaciones específicas del conjunto de datos de entrenamiento, en lugar de generalizar correctamente a nuevos datos o conjuntos de datos no vistos. **Esto suele suceder cuando el modelo es demasiado complejo** y captura el ruido en lugar de la verdadera señal subyacente (Pursky et al., 2018,p. 69)
- El algoritmo de aprendizaje automático que se utiliza comúnmente para predecir un valor numérico continuo en problemas de regresión es el "Árbol de Decisión".

Sin embargo, hay una extensión de los árboles de decisión llamada "Árboles de Regresión" o "Árboles de Regresión y Clasificación" que se utiliza específicamente para tareas de regresión. **La regresión lineal se puede utilizar para predecir valores numéricos continuos.** Por ejemplo, se puede utilizar para predecir el precio de una casa en función de su tamaño, número de habitaciones, ubicación, etc. También se puede utilizar para predecir el rendimiento académico de un estudiante en función de su nivel de inteligencia, asistencia a clase y horas de estudio (Aguilar Reyes et al., 2022,p. 77).

- El tipo de algoritmo de aprendizaje automático que se basa en la imitación del comportamiento observado para tomar decisiones se conoce como "**aprendizaje por refuerzo".** Este enfoque de aprendizaje automático implica entrenar modelos o agentes para aprender acciones y comportamientos a partir de ejemplos o demostraciones proporcionadas por expertos o fuentes de datos (Pursky et al., 2018,p. 64).
- Tanto el **oversampling como el undersampling** son técnicas utilizadas para abordar el desequilibrio de clases. Mientras que el oversampling duplica ejemplos de la clase minoritaria, el undersampling reduce ejemplos de la clase mayoritaria (*Como Lidiar Con El Desbalanceo de Datos | Alura Cursos Online*, 2024)**.**
- El **área bajo la curva ROC (AUC-ROC) es** una métrica comúnmente utilizada para evaluar la capacidad de discriminación de un modelo de clasificación. Un valor cercano a 1 indica un buen rendimiento (Pursky et al., 2018,p. 27)
- El tipo de algoritmo de aprendizaje automático que se utiliza comúnmente para problemas de recomendación de películas en plataformas como **Netflix es el "filtrado colaborativo"**. El filtrado colaborativo es una técnica de recomendación que se basa en las interacciones históricas y preferencias de

múltiples usuarios para hacer predicciones o sugerir elementos (como películas, series, productos, etc.) que un usuario específico podría disfrutar o encontrar relevante (Pursky et al., 2018,p. 21).

- Los **algoritmos de clustering**, como el DBSCAN o K-means, pueden identificar **patrones anómalos en conjuntos de datos** al agrupar datos similares y señalar valores atípicos (Pursky et al., 2018,p. 64)
- El tipo de aprendizaje automático que se utiliza para **predecir una variable objetivo basada en variables de entrada se denomina "aprendizaje supervisado".** En el aprendizaje supervisado, se entrena un modelo utilizando un conjunto de datos etiquetado que contiene pares de características (variables de entrada) y etiquetas o respuestas conocidas (variable objetivo). El objetivo es aprender una función o modelo que pueda predecir la etiqueta o respuesta correcta para nuevas entradas o datos no vistos basándose en la relación entre las variables de entrada y la variable objetivo observada en los datos de entrenamiento (Pursky et al., 2018,p. 16).

3. ÉTICA Y PRIVACIDAD EN CIENCIA DE DATOS

- El principal objetivo de las políticas de privacidad en ciencia de datos es **garantizar la protección, integridad y confidencialidad de los datos personales** y sensibles utilizados, procesados o gestionados durante proyectos de análisis, investigación y desarrollo de modelos y aplicaciones de aprendizaje automático y análisis de datos. Estas políticas buscan establecer normas, prácticas y procedimientos que aseguren el cumplimiento de los principios éticos, legales y regulatorios relacionados con la privacidad y seguridad de la información en el ámbito de la ciencia de datos (Larreta, 2023,p. 5).

El consentimiento informado en el contexto de la ciencia de datos se refiere al proceso mediante el cual los **individuos otorgan su permiso explícito y consciente para que sus datos personales sean recopilados, procesados, utilizados o compartidos en proyectos**, investigaciones y aplicaciones de análisis de datos y aprendizaje automático. Este consentimiento debe ser otorgado de manera voluntaria, clara, comprensible y específica, informando a los individuos sobre los propósitos, alcance, riesgos, beneficios y prácticas relacionadas con la utilización y gestión de sus datos personales (Larreta, 2023,p. 5)

- El principio de minimización de datos es un concepto fundamental en la protección de datos personales y privacidad que implica limitar la recopilación, procesamiento, almacenamiento y **utilización de datos personales únicamente a aquellos datos que sean estrictamente necesarios, pertinentes y proporcionales para cumplir con fines específicos** y legítimos en proyectos, actividades y operaciones que involucren la gestión y manipulación de información de individuos (Larreta, 2023,p. 11).
- El anonimato de datos **implica modificar o eliminar información personal de un conjunto de datos de manera que sea imposible identificar a individuos** específicos (Tâm et al., 2016, p. 11)
- La integridad de datos se refiere a garantizar que los **datos sean exactos, consistentes y fidedignos, evitando cualquier tipo de manipulación** o distorsión (Larreta, 2023,p. 17)
- El cumplimiento de regulaciones como el GDPR es esencial para proteger los derechos de privacidad de los individuos y **evitar posibles sanciones legales por el uso indebido o no autorizado de datos personales** (Tâm et al., 2016, p. 22)

- La ética de la transparencia se centra en **divulgar de manera clara y abierta los métodos, fuentes y procesos utilizados** en la recopilación y análisis de datos, promoviendo la confianza y la integridad en la investigación (Larreta, 2023,p. 25)
- Una práctica ética recomendada al compartir datos entre investigadores es garantizar la confidencialidad, integridad y seguridad de la información, respetando los derechos y privacidad de los participantes y sujetos de investigación, **es fundamental establecer acuerdos claros y transparentes entre investigadores** al compartir datos, garantizando la atribución adecuada, el uso ético y la protección de la propiedad intelectual (Tâm et al., 2016, p. 26)
- La responsabilidad ética en la ciencia de datos se centra en garantizar que las **decisiones y acciones basadas en datos sean justas, equitativas** y respeten los derechos y dignidad de los individuos (Larreta, 2023,p. 3)
- La ética en la comunicación de resultados i**mplica presentar la información de manera honesta, clara y comprensible**, evitando cualquier forma de manipulación o tergiversación que pueda llevar a interpretaciones erróneas (Tâm et al., 2016, p. 31)

4. BIG DATA Y COMPUTACIÓN EN LA NUBE

- El término "Volumen" en Big Data hace referencia a la **cantidad masiva de datos generados y almacenados**, siendo uno de los "tres V's" fundamentales de Big Data junto con la variedad y velocidad (Alonso-zárate et al., 2023, p. 8).
- El principal desafío asociado con la variedad de datos en Big Data es la complejidad **en la integración,** gestión, procesamiento y análisis de diferentes tipos, formatos, estructuras y fuentes de datos heterogéneas que pueden ser

voluminosas, dinámicas y dispersas en diversos sistemas, plataformas y entornos (Alonso-zárate et al., 2023, p. 12)

- Hadoop es un framework de código abierto que permite el **procesamiento distribuido de grandes conjuntos de datos** en clusters de computadoras, facilitando el almacenamiento y procesamiento de datos a gran escala (Alonso-zárate et al., 2023, p. 12)
- Para predecir eventos futuros en Big Data, **se utiliza comúnmente el análisis predictivo**. El análisis predictivo es una técnica avanzada de análisis de datos que utiliza modelos estadísticos, algoritmos de aprendizaje automático, técnicas de minería de datos y enfoques analíticos para identificar patrones, tendencias, relaciones y correlaciones en grandes volúmenes de datos y generar predicciones, estimaciones y proyecciones sobre eventos, comportamientos, resultados y tendencias futuras en diversos dominios y aplicaciones (Alonso-zárate et al., 2023, p. 28)
- En el contexto de Big Data, "velocidad" se refiere a la capacidad y **velocidad con la que los datos se generan (Frecuencia con la que se generan datos**), capturan, procesan, en tiempo real o cerca de tiempo real dentro de sistemas, plataformas y aplicaciones de Big Data (Alonso-zárate et al., 2023, p. 8)
- **Data Mining es el proceso de descubrir patrones y conocimientos útiles a partir de grandes conjuntos de datos no estructurados** utilizando técnicas como la inteligencia artificial, el aprendizaje automático y la estadística (Alonso-zárate et al., 2023, p. 32)
- Un Data Lake es un sistema de **almacenamiento que permite almacenar datos estructurados y no estructurados en su formato original**, facilitando el análisis de grandes volúmenes de datos sin requerir una estructuración previa (IBM, 2024)

- Apache Spark es un motor de **procesamiento de datos en memoria que permite el procesamiento en tiempo real de grandes conjuntos de datos**, proporcionando una velocidad significativamente mayor en comparación con otras tecnologías (Alonso-zárate et al., 2023, p. 13)
- El desafío de Big Data que se relaciona con la garantía de la precisión y fiabilidad de los datos **se denomina "veracidad".** La veracidad es uno de los "cuatro V" adicionales que se han propuesto para complementar los "tres V" tradicionales del Big Data (Volumen, Velocidad y Variedad), y se refiere a la necesidad de asegurar la calidad, precisión, fiabilidad, integridad y confiabilidad de los datos recopilados, procesados, almacenados, analizados y utilizados en proyectos, aplicaciones y operaciones de Big Data en diversos contextos y sectores (Alonso-zárate et al., 2023, p. 8)
- Una de las ventajas principales de implementar soluciones de Big Data en organizaciones es la capacidad **para reducir costos** al mejorar la eficiencia operativa, optimizar procesos y tomar decisiones basadas en datos (Alonso-zárate et al., 2023, p. 21)

5. INTELIGENCIA ARTIFICIAL

- El **aprendizaje automático** se centra en **desarrollar algoritmos que permiten a las máquinas mejorar su rendimiento** en tareas específicas a partir de la experiencia adquirida a partir de datos (Pursky et al., 2018, p. 25)
- El clustering, también conocido como análisis de agrupamiento, es una técnica de aprendizaje automático no supervisado que se utiliza principalmente **para reconocer patrones y clasificar datos en conjuntos** o grupos basados en la similitud o proximidad de características y atributos Pursky et al., 2018, p. 64)

- La regresión es un tipo de aprendizaje automático supervisado que se utiliza para predecir valores numéricos basándose en variables independientes. Se utiliza comúnmente en análisis predictivo y modelado estadístico (Pursky et al., 2018, p. 18)
- El aprendizaje por refuerzo es una técnica de IA donde un agente toma decisiones interactivamente y **aprende a través de la retroalimentación** del entorno para maximizar una recompensa acumulada (Pursky et al., 2018, p. 63)
- El procesamiento del lenguaje natural (PLN) se enfoca en permitir que las máquinas comprendan, interpreten y generen **lenguaje humano de manera eficiente,** facilitando la comunicación entre humanos y computadoras (Pursky et al., 2018, p. 44)
- Las Redes Neuronales Artificiales (RNA) están inspiradas en la estructura y función de las neuronas biológicas y se utilizan en IA para simular procesos cognitivos (**simular el comportamiento humano**) y aprender a partir de datos (Pursky et al., 2018, p. 11)

  La lógica difusa es una técnica de Inteligencia Artificial (IA) que se utiliza para manejar **la incertidumbre y la imprecisión en los datos y la toma de decisiones**. La lógica difusa permite manejar la incertidumbre y la imprecisión en los datos mediante conjuntos difusos y reglas lógicas, facilitando la toma de decisiones en entornos ambiguos (innova, 2024)
- El aprendizaje supervisado es una técnica de aprendizaje automático en la que se entrena un modelo utilizando un conjunto de datos que contiene ejemplos etiquetados. Estos ejemplos consisten en pares de entrada y salida, donde la entrada se refiere a las características o atributos y la salida es una etiqueta o valor que se desea predecir. Cuando se menciona **"predecir valores numéricos",** se

está refiriendo a problemas de regresión, que es una aplicación típica del aprendizaje supervisado (Pursky et al., 2018, p. 17)

- Las Redes Bayesianas, también conocidas como Redes de Bayes, son una técnica fundamental en la inteligencia artificial que se utiliza para modelar y representar **el razonamiento probabilístico**. Estas redes están basadas en el teorema de Bayes y permiten expresar relaciones probabilísticas entre un conjunto de variables y sus dependencias, facilitando la inferencia y el análisis de incertidumbre en sistemas y situaciones donde se requiere evaluar, calcular y tomar decisiones basadas en probabilidades, evidencias y conocimiento experto (Pursky et al., 2018, p. 68)

  Los algoritmos genéticos son una técnica de optimización y búsqueda inspirada en el proceso de evolución natural que se utiliza en inteligencia artificial. Estos algoritmos **trabajan iterativamente** para mejorar soluciones candidatas a un problema específico, aplicando operadores genéticos como selección, cruce y mutación para generar nuevas soluciones a partir de una población inicial de individuos (Tâm et al., 2016, p. 38)

Bibliografía

Aguilar Reyes, J. E., Chariguamán Maurisaca, Martínez Moscoso, M. E., & Calderón, S. H. (2022). *La estadística como una herramienta en la Metodología Científica*.

Alonso-zárate, J., Julbe, F., Lloret, M. M., & Nin, J. (2023). *Tecnologías de big data*.

Como lidiar con el desbalanceo de datos | Alura Cursos Online. (2024). https://www.aluracursos.com/blog/como-lidiar-con-el-desbalanceo-de-datos

Faraldo, P., & Pateiro, B. (2013). Tema 1. Estadística descriptiva. *Estadística y Metodología de La Investigación.*, 15. http://eio.usc.es/eipc1/BASE/BASEMASTER/FORMULARIOS-PHP-DPTO/MATERIALES/Mat_G2021103104_EstadisticaTema1.pdf

IBM. (2024). *¿Qué es un data lake?* https://mioti.es/es/que-es-un-data-lake/

innova. (2024). ▷*Lógica Difusa en Inteligencia Artificial: Una Introducción* 【*Canal Innova*】. https://canalinnova.com/logica-difusa-en-inteligencia-artificial-una-introduccion/

Larreta, H. R. (2023). *Uso Ético de los Datos*.

Orellana, L. (2001). Estadística Desctiptiva. *Estadistica Descriptiva*, 1–64. https://www.dm.uba.ar/materias/estadistica_Q/2011/1/modulo descriptiva.pdf

Pursky, O. I., Kharchenko, O. A., & Mazokha, D. P. (2018). Utilización de sistemas de recomendación basados en métodos de aprendizaje automático en la disciplina del comercio electrónico. In *New computer technology* (Vol. 16). https://doi.org/10.55056/nocote.v16i0.830

Tâm, T., Và, N. C. Ứ U., Giao, C. Ể N., Ngh, C., & Chu, Ẩ N B Ụ I. (2016). *Aspectos de seguridad en el manejo de datos sensibles*. *01*, 1–23.