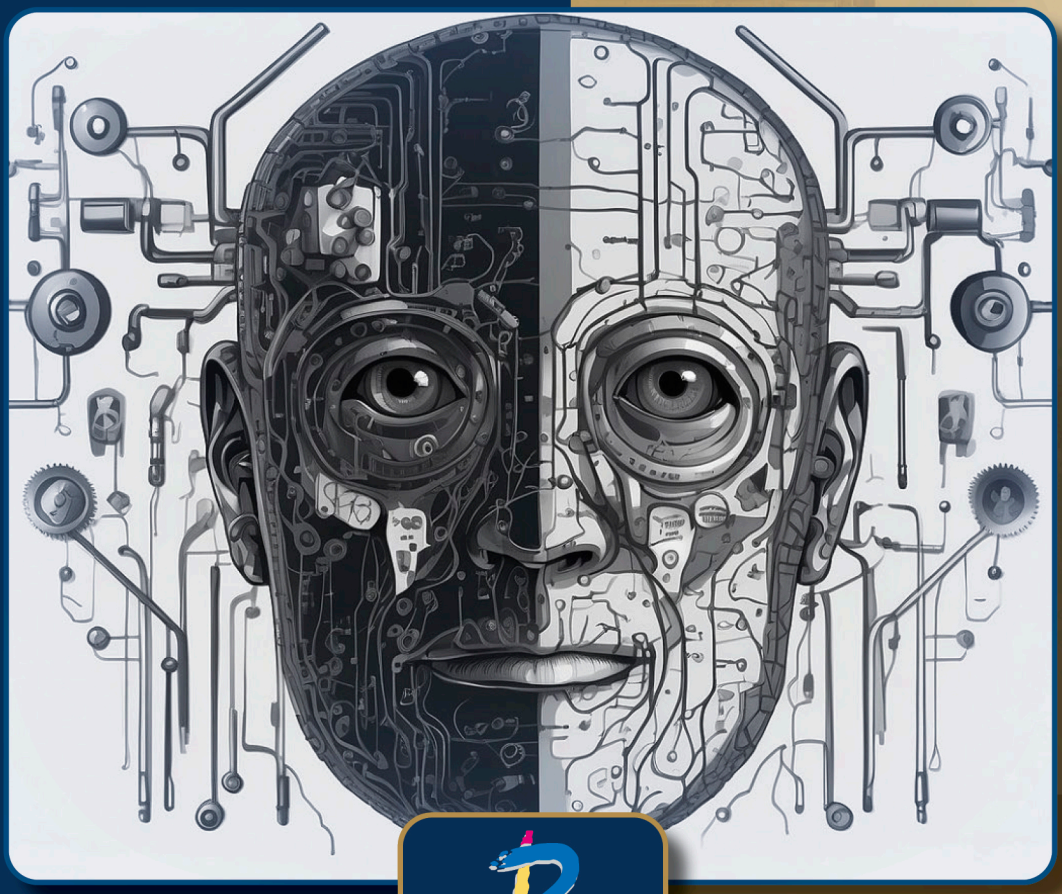
JACINTO VELASCO REBOLLEDO

MACHINE LEARNING

Fundamentos, algoritmos y aplicaciones para los negocios, industria y finanzas



DIAZ DE SANTOS

MACHINE LEARNING

Fundamentos, algoritmos y aplicaciones para los negocios, industria y finanzas

JACINTO VELASCO REBOLLEDO

MACHINE LEARNING

Fundamentos, algoritmos y aplicaciones para los negocios, industria y finanzas



Madrid • Buenos Aires • México • Bogotá

© Jacinto Velasco Rebolledo, 2024

Reservados todos los derechos.

«No está permitida la reproducción total o parcial de este libro, ni su tratamiento informático, ni la transmisión de ninguna forma o por cualquier medio, ya sea electrónico, mecánico por fotocopia, por registro u otros métodos, sin el permiso previo y por escrito de los titulares del Copyright.»

Ediciones Díaz de Santos
Internet: http//www.editdiazdesantos.com
E-mail: ediciones@editdiazdesantos.com

e-ISBN: 978-84-9052-531-9 (ebook)

ISBN: 978-84-9052-530-2 (papel)

Fotocomposición y diseño de cubiertas: P55 Servicios Culturales

ÍNDICE

Prólogo..... XI

MÓDULO I. Conceptos generales.....1

1.1. Conceptos generales de aprendizaje automático 3
1.2. Infraestructura y tipos de analítica de un proyecto de ciencia de datos.....5
1.2.1. Fundamentos sobre infraestructura de datos 5
1.2.2. Diseño de una infraestructura de datos 8
1.2.3. Tipos de analítica 13
1.3. Tipos de aprendizaje automático 15
1.3.1. Aprendizaje supervisado.....16
1.3.2. Aprendizaje no supervisado.....17
1.3.3. Aprendizaje por refuerzo.....17
1.4. Etapas de un proyecto de aprendizaje automático.....19
1.4.1. *Feature engineering*.....19
1.4.2. Etapas en un proyecto de aprendizaje automático.....20
1.5. Técnicas de optimización en aprendizaje automático.....25
1.6. Python y buenas prácticas de programación.....31
1.6.1. Conceptos fundamentales de Python 31
1.6.2. Buenas prácticas en la programación de modelos de aprendizaje automático.....41
1.7. Análisis exploratorio de datos (EDA) 42

MÓDULO II. Aprendizaje supervisado. Clasificación.....47

2.1. Técnicas de validación de modelos.....49
2.1.1. Matriz de confusión para modelos de clasificación.....49
2.1.2. Curva ROC.....53
2.1.3. Validación cruzada o *cross validation*.....57
2.1.4. Entropía e índice Gini. Curvas de aprendizaje. 59
2.1.5. Medidas de distancias 63
2.2. Teoría de la decisión: clasificación bayesiana.....65
2.2.1. Teorema de Bayes.....65
2.2.2. Algoritmo de Naive Bayes.....67
2.3. Sistemas de clasificación básicos 71
2.3.1. Árbol de decisión simple 71
2.3.2. Algoritmo KNN.....73

2.4. Sistemas de clasificación avanzados 76
2.4.1. *Random forest* 76
2.4.2. *Boosting* 80
2.5. Máquinas de soporte vectorial 87
2.6. Introducción a redes neuronales 92
2.6.1. Definición y tipos de redes neuronales 92
2.6.2. Redes neuronales multicapa 96
2.6.3. Redes neuronales convolucionales (CNN) 102
2.6.4. Redes neuronales recurrentes (RNN) 106
2.7. Procesamiento del lenguaje natural (NLP) 108
2.8. Estrategias de resolución de problemas en modelos de clasificación 115

MÓDULO III. Modelos de regresión y series temporales 119
3.1. Validación en modelos de regresión 121
3.2. Regresión lineal múltiple 128
3.3. Regresión logística 134
3.4. Modelos de árboles con regresión 139
3.5. Regresión con vector soporte 144
3.6. Series temporales con aprendizaje automático 147
3.6.1. Introducción a las series temporales y análisis de componentes 147
3.6.2. Modelos ARIMA y SARIMA 150
3.6.3. Series temporales con redes neuronales 155

MÓDULO IV. Aprendizaje no supervisado y por refuerzo 159
4.1. Análisis de componentes principales (PCA) 161
4.2. Modelos de aprendizaje no supervisado 163
4.2.1. Agrupación en clústeres: K-means 163
4.2.2. DBSCAN 167
4.2.3. Algoritmos a priori 169
4.2.4. Técnicas de validación de modelos no supervisados 173
4.3. Aprendizaje por refuerzo 177

MÓDULO V. Casos de uso 181
5.1. Caso de uso de una segmentación de prospectos 183
5.1.1. Definición del modelo 183
5.1.2. Descripción del caso de uso 184
5.1.3. Resolución técnica 184
5.1.4. Análisis de resultados 188

5.2. Caso de uso de un modelo de propensión 192
5.2.1. Definición del modelo 192
5.2.2. Descripción del caso de uso 193
5.2.3. EDA 194
5.2.4. Resolución técnica 196
5.2.5. Cálculo del tiempo de retorno 204
5.2.6. Distribución de nuevas probabilidades 208
5.2.8. Conclusiones y análisis de negocio 210
5.3. Análisis de series temporales para predecir el número de viajeros en el transporte urbano 210
5.3.1. Descripción del caso de uso 210
5.3.2 Resolución técnica usando SARIMA 211
5.3.3. Conclusiones y análisis de negocio 217
5.4. Caso de uso de un modelo de abandono (*churn*) para explicar la pérdida de clientes en el sector financiero. Optimización de parámetros 218
5.4.1. Definición del modelo 218
5.4.2. Descripción del caso de uso 219
5.4.3. EDA 220
5.4.4. Resolución técnica 223
5.4.5. Conclusiones y análisis de negocio 229
5.5. Caso de uso para el análisis de sentimiento en redes sociales mediante el teorema de Bayes. 231
5.5.1. Definición del modelo 231
5.5.2. Descripción del caso de uso 232
5.5.3. Resolución técnica 233
5.5.4. Conclusiones y análisis de negocio 235

Bibliografía 237

PRÓLOGO

En la intersección entre la innovación tecnológica y las estrategias empresariales, emerge el mundo del aprendizaje automático. En la era digital, donde los datos son valorados como el nuevo petróleo y la habilidad para tomar decisiones precisas y ágiles se considera la clave del éxito empresarial, el dominio de las herramientas y técnicas del aprendizaje automático se ha vuelto imprescindible tanto para líderes empresariales consolidados como para estudiantes universitarios.

Este campo en constante evolución no solo revoluciona la forma en que las organizaciones abordan sus operaciones y estrategias comerciales, sino que también abre un vasto espectro de oportunidades para aquellos que están dispuestos a explorarlo. Desde la personalización de experiencias para los clientes hasta la detección de patrones ocultos en grandes conjuntos de datos, el aprendizaje automático se ha convertido en un motor fundamental de la innovación y el progreso en el ámbito empresarial moderno.

En este contexto, comprender los fundamentos y aplicaciones del aprendizaje automático se convierte en una ventaja competitiva crucial. Los líderes empresariales deben estar equipados con el conocimiento necesario para aprovechar al máximo el potencial de esta tecnología, mientras que los estudiantes universitarios buscan adquirir las habilidades esenciales que les permitirán prosperar en el mercado laboral del futuro.

El autor, con más de 20 años de experiencia en la aplicación de estas metodologías para resolver casos de uso en diversas áreas empresariales, desde la financiera e industrial hasta la comercial y de marketing, ha destilado su conocimiento en este libro. La obra ha sido diseñada para ser una guía completa y accesible sobre aprendizaje automático aplicado a los negocios, representando un recurso invaluable tanto para aquellos que buscan mejorar su comprensión de las complejidades del aprendizaje automático en el contexto empresarial, como para profesores que desean impartir conocimientos actualizados y relevantes en el aula universitaria.

Dentro de las páginas de este compendio, los lectores serán guiados en un fascinante viaje que parte desde los cimientos elementales hasta las sofisticadas técnicas del aprendizaje automático, todo ello expuesto mediante un lenguaje matemático comprensible por cualquier lector con formación básica en economía, finanzas o tecnología. A lo largo de este recorrido, se explorará minuciosamente cómo aplicar algoritmos de aprendizaje supervisado y no supervisado para abordar una amplia variedad de desafíos empresariales.

Desde la anticipación de la demanda hasta la optimización de la cadena de suministro, así como el análisis de sentimientos y la personalización de ex-

periencias para los clientes, este libro se convierte en un recurso esencial para aquellos que buscan ampliar su comprensión y dominio en el ámbito del aprendizaje automático aplicado al entorno empresarial.

Este texto va más allá de meras teorías abstractas; su enfoque se concentra en proporcionar a los lectores las habilidades prácticas necesarias para afrontar los desafíos cambiantes del mundo empresarial contemporáneo. Además, está meticulosamente estructurado para ser utilizado tanto por estudiantes universitarios en busca de conocimientos fundamentales como por profesionales que aspiran a integrar el aprendizaje automático en sus actividades cotidianas.

En el libro se incluyen ejemplos en Python para ayudar a comprender mejor la teoría presentada. Estos ejemplos prácticos muestran cómo aplicar los conceptos de aprendizaje automático a situaciones empresariales concretas. No solo clarifican la teoría expuesta en el texto, sino que también proporcionan a los lectores herramientas prácticas que pueden aplicar en sus propios proyectos empresariales. Al ver cómo se implementan los conceptos de aprendizaje automático en código, los lectores pueden comprender mejor cómo adaptar y aplicar estas técnicas a sus propios casos de uso específicos dentro de sus organizaciones. Esta combinación de teoría y ejemplos prácticos en Python asegura que los lectores no solo adquieran conocimientos conceptuales sobre aprendizaje automático, sino que también desarrollen habilidades prácticas que puedan utilizar para abordar desafíos empresariales reales.

En última instancia, este libro tiene como objetivo capacitar a sus lectores para que se conviertan en líderes visionarios y tomadores de decisiones informadas, capaces de aprovechar el poder del aprendizaje automático para impulsar la innovación, la eficiencia y el crecimiento en sus organizaciones y en la sociedad en general.

Dr. Jacinto Velasco Rebolledo

Módulo I

CONCEPTOS GENERALES

1.1. CONCEPTOS GENERALES DE APRENDIZAJE AUTOMÁTICO

La inteligencia artificial (IA) es un campo de la informática que se centra en el desarrollo de sistemas y algoritmos capaces de realizar tareas que normalmente requieren inteligencia humana. La IA abarca una amplia gama de técnicas y enfoques, incluyendo el aprendizaje automático y el aprendizaje profundo.

El aprendizaje automático o *machine learning* (ML), se refiere a un subcampo de la inteligencia artificial que se centra en el estudio y desarrollo de algoritmos y modelos computacionales que permiten a los sistemas aprender y mejorar su rendimiento a partir de la experiencia, sin necesidad de una programación explícita para cada tarea específica. No hay que confundir aprendizaje automático con otros tipos de disciplinas relacionadas con la inteligencia artificial como el *deep learning* o GenAI. El primero es una subárea del aprendizaje automático que se basa en redes neuronales artificiales con múltiples capas (profundidad), lo que permite aprender representaciones de datos cada vez más abstractas y complejas. El segundo es un subcampo emergente de la inteligencia artificial que se centra en el desarrollo de modelos capaces de generar nuevos datos o contenido original, como imágenes, música o texto, que sean indistinguibles de los creados por humanos. Estos modelos utilizan técnicas avanzadas de aprendizaje profundo.

Los algoritmos de aprendizaje automático utilizan datos para identificar patrones, realizar predicciones o tomar decisiones, y se caracterizan por su capacidad para adaptarse y mejorar su desempeño a medida que se exponen a más datos.

La historia del aprendizaje automático se remonta a más de medio siglo atrás, con raíces que se extienden a los albores de la informática y la inteligencia artificial. La inteligencia artificial (IA) nació oficialmente durante la Conferencia de Dartmouth en 1956, donde se acuñó el término y se sentaron las bases para su investigación. Aunque el aprendizaje automático no se mencionó específicamente en esta conferencia, sentó las bases para el desarrollo de algoritmos de aprendizaje automático en las décadas siguientes. El psicólogo Frank Rosenblatt desarrolló el perceptrón, uno de los primeros algoritmos de aprendizaje supervisado. Este modelo de red neuronal simple fue capaz de aprender a reconocer patrones en los datos de entrada y se utilizó en una variedad de aplicaciones tempranas de reconocimiento de patrones. Durante estos años, se desarrollaron varias técnicas de aprendizaje automático simbólico, como el árbol de decisiones y los sistemas expertos. También se introdujo la lógica difusa,

una técnica para modelar la incertidumbre en los datos, que se utiliza en aplicaciones de control y reconocimiento de patrones. Durante la década de 1990, surgieron importantes avances en algoritmos de aprendizaje supervisado, como la regresión logística y las máquinas de soporte vectorial (SVM). Estos modelos fueron ampliamente utilizados en aplicaciones de clasificación y regresión en una variedad de campos. Con el auge del *big data* en la década de 2000, el aprendizaje automático experimentó un renacimiento. La disponibilidad de grandes conjuntos de datos y avances en hardware permitieron el desarrollo de modelos más complejos y sofisticados, como las redes neuronales profundas (*deep learning*). Estos modelos han demostrado un rendimiento excepcional en tareas de percepción, como reconocimiento de imágenes y procesamiento de lenguaje natural. En los últimos años, el aprendizaje automático se ha vuelto omnipresente en la vida cotidiana, con aplicaciones que van desde motores de búsqueda y recomendaciones de productos hasta reconocimiento facial y sistemas de conducción autónoma. En resumen, la historia del aprendizaje automático es una narrativa de continuo progreso y evolución, impulsada por avances en la teoría, el hardware y la disponibilidad de datos. Desde sus modestos comienzos en los laboratorios de investigación hasta su integración en nuestra vida diaria, el aprendizaje automático continúa transformando la forma en que interactuamos con la tecnología y el mundo que nos rodea.

La inteligencia artificial (IA) está transformando rápidamente la forma en que las organizaciones operan y toman decisiones. La gestión del cambio se vuelve crucial en este proceso de adopción de la IA, ya que implica modificar las estructuras organizativas, los procesos y las culturas para aprovechar al máximo el potencial de esta tecnología innovadora. La implementación de la IA puede generar resistencia y temores entre los empleados debido a la incertidumbre sobre cómo afectará sus roles y responsabilidades. Por lo tanto, es fundamental comunicar de manera efectiva los beneficios y objetivos de la IA y proporcionar capacitación adecuada para garantizar una transición suave y exitosa.

La gestión del cambio también implica establecer una visión clara y un liderazgo sólido para guiar a la organización a través de la transformación. Esto incluye la identificación de los impulsores clave del cambio, la alineación de los procesos con los objetivos estratégicos y la promoción de una cultura de innovación y aprendizaje continuo.

Además, es importante involucrar a los empleados en el proceso de adopción de la IA, fomentando la participación, la colaboración y la retroalimentación para impulsar la aceptación y el compromiso. Esto puede incluir la creación de grupos de trabajo interdisciplinarios y la promoción de la experimentación y la creatividad.

El futuro del aprendizaje automático (ML) es un horizonte prometedor lleno de innovación y avances que están transformando rápidamente múltiples industrias y aspectos de la vida cotidiana. Una de las principales tendencias que

se espera en el futuro del ML es la evolución hacia modelos más avanzados y complejos, alimentados por el acceso a grandes volúmenes de datos y el aumento de la potencia informática. Estos modelos podrán realizar tareas más sofisticadas, como el procesamiento de lenguaje natural, la visión por computadora y la toma de decisiones más autónoma.

Además, se espera que el aprendizaje automático se integre aún más en nuestra vida diaria a través de dispositivos inteligentes y sistemas de asistencia personalizados. Esto incluye avances en la automatización del hogar, la conducción autónoma, la atención médica personalizada y la recomendación de contenido en plataformas de entretenimiento y comercio electrónico. Estos avances mejorarán la eficiencia, la comodidad y la calidad de vida para las personas en todo el mundo.

Otra tendencia importante es la democratización del aprendizaje automático, que implica hacer que esta tecnología sea más accesible para una gama más amplia de personas y organizaciones. Esto incluye el desarrollo de herramientas y plataformas de ML de código abierto, así como la simplificación de los procesos de desarrollo y despliegue de modelos. La democratización del ML permitirá a más personas y empresas aprovechar el poder de esta tecnología para resolver problemas y generar valor.

En el ámbito empresarial, el aprendizaje automático continuará desempeñando un papel crucial en la optimización de procesos, la toma de decisiones basada en datos y la personalización de productos y servicios. Se espera que las empresas inviertan cada vez más en capacidades de ML y desarrollen sus propios equipos internos de ciencia de datos para aprovechar al máximo estas tecnologías.

Por último, el futuro del aprendizaje automático también plantea desafíos importantes, como la ética y la responsabilidad en el uso de datos y algoritmos, así como preocupaciones sobre la privacidad y la seguridad de la información. A medida que el ML continúe avanzando, será fundamental abordar estos desafíos de manera proactiva y desarrollar marcos éticos y regulatorios sólidos para garantizar un uso responsable y beneficioso de esta tecnología. En resumen, el futuro del aprendizaje automático es emocionante y lleno de posibilidades, pero también requiere un enfoque cuidadoso y reflexivo para garantizar que se utilice de manera ética y responsable para el beneficio de la sociedad.

1.2. INFRAESTRUCTURA Y TIPOS DE ANALÍTICA DE UN PROYECTO DE CIENCIA DE DATOS

1.2.1. Fundamentos sobre infraestructura de datos

Establecer una infraestructura de datos robusta y eficiente es fundamental para la implementación exitosa de proyectos de aprendizaje automático. En todos los procesos de datos se pueden definir las siguientes etapas:



Figura 1.1. Etapas de un proyecto de datos.

En el contexto de la gestión de datos, la infraestructura y la arquitectura desempeñan un papel fundamental en la creación de sistemas sólidos y eficientes para el manejo, procesamiento y análisis de datos. La infraestructura de datos se refiere al conjunto de tecnologías, herramientas y procesos utilizados para gestionar y almacenar datos de manera eficiente y segura, mientras que la arquitectura de datos se centra en el diseño y la organización de estos componentes para satisfacer las necesidades específicas del negocio.

La infraestructura de datos abarca una variedad de componentes, incluidos sistemas de almacenamiento de datos, bases de datos, herramientas de procesamiento y análisis de datos, servidores, redes y software de gestión de datos. Estos componentes trabajan juntos para garantizar que los datos estén disponibles, accesibles y seguros para su uso en diferentes aplicaciones y procesos empresariales. La infraestructura de datos también implica la implementación de prácticas de seguridad, respaldo y recuperación de datos para proteger la integridad y confidencialidad de la información.

Por otro lado, la arquitectura de datos se centra en el diseño y la organización de estos componentes para satisfacer las necesidades específicas del negocio. Esto implica definir la estructura de los datos, establecer estándares y normas para su almacenamiento y organización, y diseñar flujos de datos eficientes que permitan la integración y el intercambio de información entre diferentes sistemas y aplicaciones. La arquitectura de datos también incluye la definición de modelos de datos, esquemas de bases de datos y metadatos que proporcionan un marco para la comprensión y el uso de los datos en toda la organización.

La infraestructura y la arquitectura de datos son elementos clave en la gestión de datos, proporcionando la base tecnológica y organizativa necesaria para garantizar que los datos sean gestionados de manera eficiente, segura y con-

forme a las necesidades del negocio. Un enfoque sólido en la infraestructura y la arquitectura de datos ayuda a las organizaciones a maximizar el valor de sus datos y a aprovechar todo su potencial para la toma de decisiones y la innovación empresarial.

A continuación, presentamos un glosario con definiciones de los términos clave relacionados con la infraestructura y la arquitectura de datos, lo que te ayudará a comprender mejor estos conceptos y cómo se interrelacionan entre sí.

- **Arquitectura de datos.** Diseño estructurado que define cómo se almacenan, organizan, acceden y gestionan los datos en una organización.
- **Lago de datos o *data lake*.** Repositorio centralizado de almacenamiento de datos que permite almacenar una amplia variedad de datos en su formato original, sin necesidad de estructura previa.
- **Almacén de datos o *data warehouse*.** Base de datos centralizada y organizada diseñada para almacenar y gestionar grandes volúmenes de datos estructurados, optimizados para consultas analíticas y generación de informes.
- **ETL (*extract, transform, load*).** Proceso que implica la extracción de datos desde una o varias fuentes, la transformación de los datos para que se ajusten a un formato común y su carga en un destino, como un *data warehouse.*
- ***Data mart*.** Subconjunto de un *data warehouse* que contiene datos específicos de un área funcional o departamento de la organización, optimizado para un conjunto particular de usuarios y casos de uso.
- **Almacén operacional (*operational data store*, ODS).** Base de datos centralizada que integra datos de varias fuentes operativas en tiempo real para proporcionar una vista única y actualizada de los datos de la organización.
- **API (interfaz de programación de aplicaciones).** Conjunto de reglas y protocolos que permiten a diferentes sistemas de software comunicarse entre sí y compartir datos y funcionalidades.
- **Arquitectura orientada a servicios (*service-oriented architecture*, SOA).** Enfoque arquitectónico que descompone una aplicación en servicios independientes y reutilizables, que se pueden acceder y combinar para cumplir con los requisitos comerciales.
- **Base de datos relacional.** Una base de datos relacional es un tipo de base de datos que organiza los datos en tablas con filas y columnas, y utiliza claves primarias y claves foráneas para establecer relaciones entre las tablas. Está basada en el modelo relacional propuesto por Edgar Codd en la década de 1970 y utiliza el lenguaje SQL (*structured query language*) para realizar consultas y manipulaciones de datos. Ejemplos incluyen MySQL, PostgreSQL y Oracle Database.

- **Base de datos no relacional (NoSQL).** Una base de datos no relacional, también conocida como NoSQL, es un tipo de base de datos diseñada para manejar grandes volúmenes de datos no estructurados o semiestructurados de manera flexible y escalable. A diferencia de las bases de datos relacionales, no tienen un esquema fijo y pueden almacenar datos en diferentes formatos, como documentos, grafos o pares clave-valor. Ejemplos incluyen MongoDB, Cassandra y Redis.
- **Tecnologías *cloud*.** Estas tecnologías proporcionan recursos informáticos y de almacenamiento bajo demanda a través de internet. Permiten escalar recursos según las necesidades del análisis de datos, ofreciendo flexibilidad y agilidad.
- **Bases de datos en tiempo real.** Estas bases de datos están diseñadas para manejar flujos continuos de datos y proporcionar resultados instantáneos a medida que los datos llegan. Son ideales para aplicaciones que requieren respuestas en tiempo real, como sistemas de monitoreo, análisis de transmisiones en vivo y detección de anomalías. Ejemplos incluyen Apache Kafka y similares.
- **Bases de datos en *batch*.** Por otro lado, las bases de datos en batch procesan grandes volúmenes de datos en lotes, generalmente en intervalos programados. Son adecuadas para aplicaciones que no requieren respuestas instantáneas y pueden tolerar cierto retraso en el procesamiento, como análisis de datos históricos y generación de informes.
- ***Edge computing*.** Es un modelo distribuido que consiste en procesar y almacenar datos cerca de su origen, en el borde de la red, en lugar de enviarlos a una ubicación centralizada, como un centro de datos o la nube, para su procesamiento y almacenamiento. Este es muy usado para procesos de IoT (*internet of things*).

1.2.2. Diseño de una infraestructura de datos

En el ámbito del aprendizaje automático, una infraestructura de datos bien diseñada es fundamental para el éxito de los proyectos. Aquí te presentamos algunas prácticas clave que te ayudarán a establecer una infraestructura robusta y eficiente para el manejo de datos en tus iniciativas de aprendizaje automático. Estas prácticas abarcan desde la recopilación y almacenamiento de datos de alta calidad hasta la implementación de pipelines de datos automatizados, la gestión de versiones y el mantenimiento continuo. A través de estas mejores prácticas, podrás garantizar la integridad, la precisión y la escalabilidad de tu infraestructura de datos, preparándola para satisfacer las demandas cambiantes del negocio y del análisis de datos en el mundo del aprendizaje automático.

Recopilación y almacenamiento de datos de alta calidad

Es fundamental garantizar la recopilación de datos de alta calidad que sean pertinentes y representativos para el problema específico que estás tratando de abordar. Esto implica no solo la obtención de datos, sino también la selección cuidadosa de aquellos que realmente son relevantes para tu análisis o investigación.

Además, es crucial emplear sistemas de almacenamiento de datos escalables y robustos que puedan manejar eficientemente grandes volúmenes de información. Estos sistemas deben ser capaces de adaptarse a medida que aumenta la cantidad de datos, asegurando que la información esté disponible y accesible cuando sea necesaria, sin comprometer la integridad o la seguridad de los datos. La capacidad de escalar vertical u horizontalmente según sea necesario es esencial para mantener el rendimiento y la confiabilidad del sistema a medida que crece la cantidad de datos almacenados.

Al implementar estos sistemas, es importante considerar no solo la capacidad de almacenamiento, sino también la velocidad de acceso a los datos, la redundancia para garantizar la disponibilidad continua y los mecanismos de seguridad para proteger la información confidencial. Esto garantizará que los datos estén disponibles y sean confiables para respaldar tus análisis y decisiones.

Limpieza y preprocesamiento de datos

Es crucial llevar a cabo una limpieza exhaustiva de los datos antes de utilizarlos en modelos de aprendizaje automático para garantizar la precisión y la fiabilidad de los resultados. Esto implica identificar y eliminar valores atípicos, manejar datos faltantes y corregir errores que puedan afectar la calidad de los datos y, en última instancia, los resultados del modelo. Para abordar estos aspectos, se deben seguir varios pasos:

- **Identificar y eliminar valores atípicos**. Los valores atípicos pueden distorsionar los resultados del modelo y deben ser tratados adecuadamente. Esto puede implicar eliminarlos si se consideran errores de entrada o si están claramente fuera del rango esperado, o puede requerir técnicas más avanzadas como la transformación de datos o el uso de modelos robustos que sean menos sensibles a los valores atípicos.
- **Manejar datos faltantes.** Los datos faltantes son comunes en conjuntos de datos reales y deben ser abordados de manera cuidadosa. Dependiendo del volumen de datos faltantes y la naturaleza de los datos, se pueden emplear diversas técnicas, como la imputación de valores utilizando la media, la mediana o técnicas más avanzadas como la regresión. En algunos casos, puede ser apropiado descartar las muestras con datos faltantes

si representan una proporción pequeña del conjunto de datos y no afectan significativamente su representatividad.
- **Corregir errores.** Es importante identificar y corregir cualquier error en los datos, ya sea inconsistencias en la estructura de los datos, errores de entrada o cualquier otra anomalía que pueda afectar la calidad de los resultados del modelo.
- **Preprocesamiento de datos.** Una vez que se han abordado los problemas de calidad de los datos, es necesario realizar el preprocesamiento necesario para preparar los datos para su uso en modelos de aprendizaje automático. Esto puede incluir la normalización de datos numéricos para que todas las características estén en la misma escala, lo que ayuda a evitar que las características con valores más grandes dominen el modelo. Además, las variables categóricas suelen necesitar ser codificadas en un formato numérico adecuado para su uso en modelos de aprendizaje automático, como la codificación *one-hot*.

Diseño de pipelines de datos

La implementación de pipelines de datos es fundamental para garantizar la eficiencia, consistencia y reproducibilidad en el flujo de datos dentro de un proyecto de aprendizaje automático. Estos pipelines son sistemas automatizados que gestionan el flujo de datos desde su ingestión inicial hasta su preparación y transformación, y finalmente su entrega al modelo de aprendizaje automático. Al implementar pipelines de datos, se logra una serie de beneficios clave:

- **Eficiencia.** Los pipelines automatizan tareas repetitivas y garantizan un flujo de datos fluido y eficiente. Esto libera a los científicos de datos y desarrolladores de la carga de trabajo manual, permitiéndoles centrarse en tareas más complejas y estratégicas.
- **Consistencia.** Al estandarizar el proceso de preparación de datos a través de un pipeline, se asegura que los datos sean tratados de manera coherente en cada ejecución. Esto ayuda a minimizar errores y asegura la consistencia de los resultados.
- **Reproducibilidad.** Al encapsular todo el proceso de preparación de datos en un pipeline, se facilita enormemente la reproducibilidad de los resultados. Cada vez que se ejecuta el pipeline con los mismos datos de entrada, se obtendrán los mismos resultados de salida, lo que facilita la validación y la colaboración en el proyecto.
- **Escalabilidad.** Los pipelines de datos pueden escalar fácilmente para manejar grandes volúmenes de datos. Esto es especialmente importante en entornos donde los conjuntos de datos pueden crecer con el tiempo, ya que los pipelines pueden adaptarse para manejar este crecimiento sin comprometer la eficiencia o la calidad del proceso.

- **Mantenibilidad.** Al encapsular la lógica de preparación de datos en un pipeline, se simplifica enormemente el mantenimiento del código. Los cambios en el proceso de preparación de datos pueden realizarse de manera centralizada en el pipeline, lo que facilita la gestión de versiones y la colaboración entre equipos.

Gestión de versiones de datos

Implementar un sistema de gestión de versiones de datos es esencial para rastrear y controlar los cambios en los conjuntos de datos a lo largo del tiempo. Esto permite una supervisión detallada de las modificaciones realizadas, desde la ingestión inicial de los datos hasta cualquier transformación o limpieza realizada durante el proceso de preparación para el modelado. Al seleccionar una herramienta adecuada, como DVC (*data version control*) o *Pachyderm*, se establece un repositorio de datos donde se almacenan todas las versiones del conjunto de datos junto con metadatos relevantes. Cada vez que se realiza una modificación en los datos, se crea una nueva versión en este repositorio, lo que proporciona un historial completo de cambios que pueden ser rastreados y revertidos si es necesario.

La documentación de los cambios es una parte crucial de este proceso, ya que proporciona contexto sobre las modificaciones realizadas, quién las realizó y con qué propósito. Esto facilita la comprensión de los cambios realizados y ayuda a mantener la coherencia y la transparencia en el desarrollo de modelos de aprendizaje automático. Además, la colaboración efectiva se ve facilitada mediante la asignación de roles y permisos en el repositorio de datos, lo que permite a los miembros del equipo trabajar de manera conjunta de manera segura y organizada

En caso de que surja la necesidad de revertir a una versión anterior del conjunto de datos, el sistema de gestión de versiones de datos proporciona la capacidad de hacerlo de manera rápida y eficiente. Esto puede ser útil en situaciones donde se introducen errores durante la preparación de los datos o cuando se necesita comparar el rendimiento de un modelo utilizando diferentes versiones del conjunto de datos. En resumen, la implementación de un sistema de gestión de versiones de datos garantiza la integridad, la trazabilidad y la reproducibilidad en el manejo de datos en proyectos de aprendizaje automático y análisis de datos.

Monitorización y mantenimiento

Es fundamental establecer sistemas de monitorización para supervisar el rendimiento de los modelos en producción y detectar posibles problemas o degradaciones en su desempeño. Estos sistemas de monitorización pueden incluir métricas clave como la precisión, la precisión del modelo, la tasa de error, el

tiempo de respuesta y otros indicadores relevantes para evaluar el rendimiento del modelo en tiempo real.

Además, es importante implementar alertas automatizadas que se activen cuando se detecten anomalías o desviaciones significativas en las métricas de rendimiento del modelo. Estas alertas pueden notificar al equipo responsable para que puedan investigar y abordar cualquier problema potencial lo antes posible, minimizando así cualquier impacto en la calidad del servicio.

Para garantizar un funcionamiento óptimo y la integridad de los datos, se debe realizar un mantenimiento regular de la infraestructura de datos. Esto implica monitorear y gestionar la capacidad de almacenamiento, asegurar la disponibilidad y la redundancia de los sistemas de almacenamiento de datos, y realizar copias de seguridad regulares para proteger contra la pérdida de datos.

Además, se deben implementar procesos de limpieza y saneamiento de datos periódicos para eliminar datos obsoletos, duplicados o irrelevantes, y garantizar que los conjuntos de datos utilizados por los modelos de aprendizaje automático sean precisos y actualizados.

El mantenimiento regular también implica la actualización y optimización continua de los modelos en producción, ya sea para mejorar su rendimiento, adaptarse a cambios en los datos de entrada o en el entorno operativo, o para incorporar nuevos conocimientos y técnicas.

Escalabilidad y flexibilidad

Al diseñar la infraestructura de datos, es crucial asegurarse de que sea escalable y pueda crecer para satisfacer las necesidades cambiantes del negocio y el aumento en el volumen de datos. Para lograr esto, se pueden implementar varias estrategias:

- **Arquitectura modular.** Diseña la infraestructura de datos utilizando una arquitectura modular que permita agregar o modificar componentes fácilmente a medida que crece la demanda. Esto facilita la escalabilidad al permitir la expansión de la infraestructura de manera incremental según sea necesario.
- **Almacenamiento escalable.** Utiliza sistemas de almacenamiento escalables que puedan manejar grandes volúmenes de datos de manera eficiente, como sistemas de almacenamiento en la nube o bases de datos distribuidas. Estos sistemas deben ser capaces de aumentar su capacidad de almacenamiento según sea necesario sin comprometer el rendimiento.
- **Procesamiento distribuido.** Emplea técnicas de procesamiento distribuido para distribuir la carga de trabajo entre múltiples nodos o servidores. Esto permite escalar horizontalmente al agregar más recursos de

procesamiento según sea necesario para manejar grandes volúmenes de datos y cargas de trabajo intensivas en procesamiento.

- **Automatización.** Implementa procesos de automatización para administrar y escalar la infraestructura de datos de manera eficiente. Esto incluye la automatización de tareas como aprovisionamiento de recursos, monitoreo de rendimiento y ajuste de capacidad para garantizar que la infraestructura pueda adaptarse dinámicamente a las fluctuaciones en la demanda.
- **Flexibilidad en la compatibilidad de datos.** Diseña la infraestructura de datos para que sea compatible con una amplia variedad de tipos de datos y modelos de aprendizaje automático. Esto implica utilizar formatos de datos estándar y herramientas que puedan manejar datos estructurados y no estructurados, así como integrar diferentes tipos de modelos de aprendizaje automático según sea necesario para abordar diversos casos de uso.
- **API y servicios orientados a microservicios.** Utiliza API y servicios orientados a microservicios para crear una infraestructura flexible y modular que pueda adaptarse rápidamente a las necesidades cambiantes del negocio. Esto permite la integración fluida de nuevos servicios y funcionalidades sin afectar la integridad de la infraestructura existente.

1.2.3. Tipos de analítica

El análisis de datos desempeña un papel crucial en la era moderna, donde la abundancia de información digital ofrece oportunidades sin precedentes para comprender y tomar decisiones informadas en una amplia gama de campos. Dentro de este ámbito, existen tres tipos principales de análisis de datos: descriptivo, predictivo y prescriptivo, cada uno con su enfoque distintivo y su conjunto de técnicas.

Análisis descriptivo

El análisis descriptivo es una fase crucial en cualquier proyecto de análisis de datos o modelado predictivo. Su objetivo principal es explorar y comprender los datos disponibles de manera integral. En esta etapa, se emplean herramientas como estadísticas descriptivas, gráficos y tablas para resumir tendencias, patrones y distribuciones presentes en los datos. Este enfoque ofrece una visión retrospectiva, permitiendo comprender qué ha sucedido en el pasado.

Al emplear estadísticas descriptivas, se calculan medidas como la media, la mediana, la moda, la desviación estándar, entre otras, para comprender la centralidad, la dispersión y la forma de la distribución de los datos. Estas medidas proporcionan una idea general de la distribución y variabilidad de los datos.

Los gráficos descriptivos son herramientas visuales poderosas que permiten representar los datos de manera intuitiva. Gráficos como histogramas, diagra-