OBS Business School

RED DE EDUCACIÓN SUPERIOR
Planeta Formación y Universidades

Raúl Gómez Martínez

Big data, data science & artificial intelligence

Material de Estudio:
3. Data Science: principales conceptos y aplicaciones.

Máster de Formación Permanente en Machine Learning e Inteligencia Artificial

Índice

Data Science: principales conceptos y aplicaciones.

- Introducción a la Ciencia de Datos
  - Definición y contexto histórico
  - Importancia y evolución en el panorama actual
- Fundamentos y Métodos de la Ciencia de Datos
  - Rol del Data Scientist y habilidades requeridas
  - Metodologías de Análisis:
    - Análisis Exploratorio de Datos (EDA)
    - La Metodología CRISP DM
  - Modelado y Predicción
- Estudios de Caso y Ejemplos Prácticos

Profesor:

Dr. Raúl Gómez Martínez

Profesor en Finanzas en la Universidad Rey Juan Carlos
Codirector del Máster en Asesoramiento y Planificación Financiera
Socio fundador de InvestMood Fintech
Consejero delegado de Open 4 Blockchain

Introducción a la Ciencia de Datos

La Ciencia de Datos es el campo multidisciplinario que se enfoca en la extracción de conocimiento y perspicacia a partir de grandes volúmenes de datos, utilizando una combinación de métodos científicos, herramientas analíticas y habilidades técnicas. Esta disciplina abarca una diversidad de áreas, incluyendo estadística, matemáticas, programación, visualización de datos y dominios específicos de aplicación.
En el corazón de la Ciencia de Datos yace la capacidad para recolectar, almacenar y analizar información compleja proveniente de diversas fuentes, lo cual permite descubrir patrones, tendencias y relaciones significativas. Este enfoque va más allá de la mera recopilación de datos, buscando transformarlos en entendimiento útil y accionable.

Los científicos de datos emplean una variedad de técnicas, como el análisis estadístico, el aprendizaje automático y la minería de datos, para explorar conjuntos masivos de información. Combinan habilidades de programación para manipular datos con herramientas especializadas para visualizarlos de manera comprensible.

Esta disciplina es fundamental en numerosos campos, desde la medicina y las finanzas hasta el comercio minorista y la investigación científica. Permite la toma de decisiones informadas, la identificación de oportunidades de negocio, el diseño de estrategias predictivas y la automatización de procesos, todo basado en la comprensión profunda de los datos.
En resumen, la Ciencia de Datos es el puente entre la abrumadora cantidad de datos generados cada día y el conocimiento significativo que puede derivarse de ellos. Su capacidad para convertir datos en ideas valiosas tiene un impacto revolucionario en la forma en que las organizaciones y sociedades toman decisiones, innovan y resuelven problemas complejos.

Definición y contexto histórico

La Ciencia de Datos es un campo interdisciplinario que se enfoca en extraer conocimiento y perspectivas significativas a partir de conjuntos de datos complejos. Esta disciplina combina elementos de estadística, matemáticas, ciencias de la computación y dominios específicos de aplicación para analizar información y obtener insights que impulsen la toma de decisiones.

Su contexto histórico se remonta a décadas atrás, aunque su explosión como disciplina independiente se dio a principios del siglo XXI. Algunos hitos históricos relevantes incluyen:

- (1960-1970) **Emergencia de la Estadística Computacional:** Se comenzaron a utilizar computadoras para realizar análisis estadísticos más complejos, sentando las bases para el procesamiento de datos a gran escala.
- (1980-1990) **Desarrollo de la Minería de Datos:** Surgió el interés en la exploración de grandes conjuntos de datos para descubrir patrones y tendencias, impulsando la evolución de la minería de datos.
- (2000s) **Explosión de Datos:** El aumento exponencial en la generación de datos, especialmente a través de internet, creó la necesidad de métodos y herramientas para manejar y analizar grandes volúmenes de información.

La definición formal de Ciencia de Datos se cristalizó cuando el término comenzó a ganar popularidad en la década de 2010. Se define como un campo que utiliza técnicas y teorías de estadística, aprendizaje automático, análisis de datos y computación para analizar y comprender fenómenos basados en datos.

Su definición ha evolucionado para reflejar su enfoque interdisciplinario, que abarca desde la recopilación y limpieza de datos hasta su análisis, modelado predictivo y visualización, con el objetivo de generar información significativa y práctica para la toma de decisiones.

El crecimiento explosivo de la Ciencia de Datos se debe en gran medida a la accesibilidad a grandes conjuntos de datos, el avance de la tecnología computacional y el desarrollo de algoritmos más sofisticados para analizar información compleja.

Esta disciplina se ha convertido en un pilar fundamental en numerosos sectores, desde la industria hasta la investigación académica, y su continua evolución sigue moldeando la forma en que utilizamos y entendemos los datos en la actualidad.

Importancia y evolución en el panorama actual

La importancia y la evolución de la Ciencia de Datos en el panorama actual son fenomenales y están transformando fundamentalmente la manera en que se toman decisiones, se innova y se abordan los desafíos en diversas industrias.

La Ciencia de Datos proporciona información precisa y accionable, permitiendo decisiones más informadas y estratégicas. Facilita la identificación de oportunidades de negocio y el desarrollo de nuevos productos y servicios basados en análisis profundos de datos, ya que permite prever tendencias futuras y riesgos potenciales, desde el comportamiento del consumidor hasta riesgos financieros, facilitando medidas preventivas. Para ello, utiliza datos para personalizar productos, servicios y experiencias, mejorando la satisfacción del cliente.

La ciencia de datos se apoya en avances en algoritmos y hardware han ampliado las capacidades de análisis y procesamiento de datos. Otro factor importante es la generación de materia prima, la generación masiva de datos provenientes de diversas fuentes impulsa la necesidad de técnicas más sofisticadas para su

análisis. Por todo ello, se han desarrollado y perfeccionado herramientas específicas para la gestión, análisis y visualización de grandes volúmenes de datos, fruto de la interdisciplinariedad que le caracteriza. La Ciencia de Datos se fusiona con áreas como la IA, el IoT, la nube y la ciberseguridad, ampliando sus aplicaciones y capacidades.

La Ciencia de Datos es el motor impulsor detrás de la transformación digital en empresas y organizaciones. Desde la automatización de procesos hasta la toma de decisiones basadas en datos, su papel es crucial en la adaptación al entorno digital actual y futuro.

Fundamentos y Métodos de la Ciencia de Datos

Los fundamentos y métodos de la Ciencia de Datos se basan en una combinación de disciplinas que incluyen estadística, matemáticas, programación, conocimientos de dominio específico y técnicas de visualización de datos (Aguilar, 2016).

La ciencia de datos se fundamenta en la estadística, empleando métodos para resumir y describir conjuntos de datos y utilizando muestras de datos para hacer inferencias sobre poblaciones más amplias. Las matemáticas aportan los fundamentos necesarios para entender y manipular datos en espacios multidimensionales, utilizando **h**erramientas para comprender y derivar algoritmos y modelos matemáticos.

Además de los conocimientos matemáticos y estadísticos es imprescindible el conocimiento sobre el área de aplicación específica para interpretar adecuadamente los resultados y tomar decisiones informadas.

A partir de hay se debe trabajar con la materia prima, los datos, asumiendo la importancia del preprocesamiento de datos, identificando y corrigiendo errores, valores atípicos y datos faltantes, así como transformando los datos con normalización y codificación para su análisis. En este preprocesamiento de datos es muy útil el Análisis Exploratorio de Datos (EDA) que emplea gráficos y representaciones visuales para entender la estructura y patrones en los datos, así como la relación entre variables para identificar correlaciones y tendencias.

Todo ello nos lleva al modelado predictivo que utiliza algoritmos para construir modelos predictivos basados en datos históricos. Estos modelos deben ser evaluados para garantizar su precisión y generalización.

Finalmente, el científico de datos debe tener capacidad para comunicar hallazgos de manera clara y efectiva a audiencias no técnicas, y traducir la relevancia de esos hallazgos técnicos en información comprensible para la toma de decisiones.

Estos fundamentos y métodos forman la base de la Ciencia de Datos, proporcionando las herramientas necesarias para recopilar, procesar, analizar y comunicar información valiosa a partir de conjuntos de datos complejos. La combinación de estas habilidades y conocimientos es fundamental para el éxito en el campo de la Ciencia de Datos.

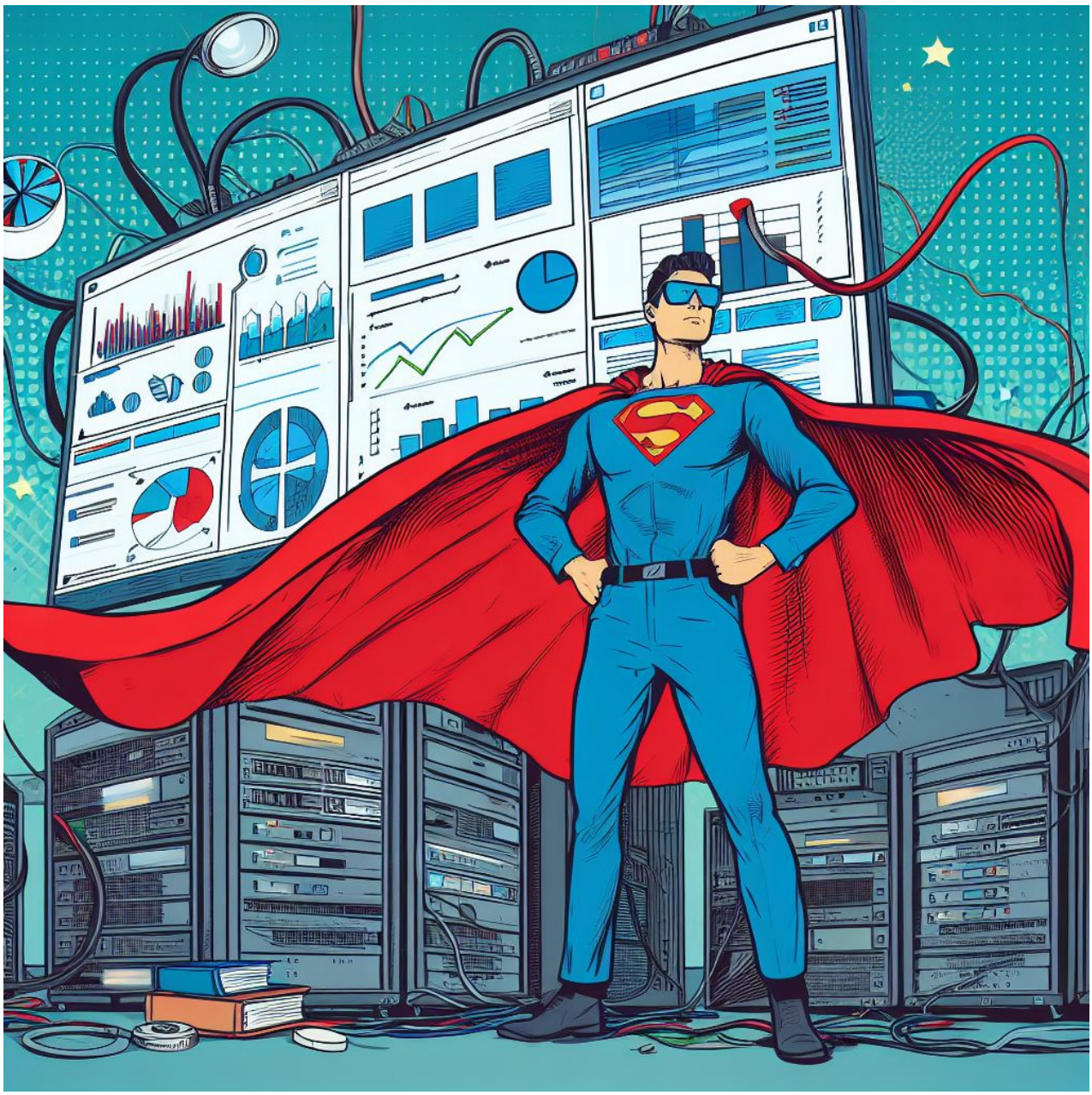
Rol del Data Scientist y habilidades requeridas

El rol del Data Scientist es crucial en la era actual, ya que se encarga de extraer información valiosa a partir de datos complejos para impulsar la toma de decisiones estratégicas. Este rol requiere una combinación única de habilidades técnicas, conocimientos de dominio y capacidades analíticas. Aquí están las habilidades clave y el rol que desempeña un Data Scientist:

Las responsabilidades del Data Scientist abarcan:

- **Comprensión del Contexto Empresarial**: Comprender el entorno empresarial y los problemas específicos del sector para aplicar soluciones de datos efectivas.

- **Procesamiento y Análisis de Datos**: Habilidad para limpiar, transformar y preprocesar datos para su análisis, y explorar y visualizar datos para identificar patrones y tendencias relevantes.
- **Modelado y Predicción**: Construir y validar modelos predictivos para tomar decisiones basadas en datos. Mejorar y ajustar algoritmos para mejorar la precisión y eficiencia de los modelos.
- **Comunicación y Presentación**: Capacidad para traducir resultados técnicos en información comprensible para diferentes audiencias.



Para asumir estas responsabilidades en necesario acumular las siguientes habilidades:

- **Conocimientos Técnicos**: Dominio de lenguajes como Python, R, SQL para manipular datos y construir modelos, y experiencia con bibliotecas de aprendizaje automático (scikit-learn, TensorFlow, etc.) y herramientas de visualización (Matplotlib, Tableau, etc.).
- **Estadística y Matemáticas**: Conocimientos sólidos en estadística para inferencia, pruebas de hipótesis y análisis multivariado, así como comprender conceptos de álgebra lineal, cálculo y teoría de la probabilidad.

- **Pensamiento Analítico y Resolución de Problemas**: Pensamiento crítico para abordar problemas complejos y encontrar soluciones innovadoras basadas en datos, sin olvidar la habilidad para descomponer problemas complejos en partes manejables y desarrollar estrategias efectivas de resolución.

Además, el científico de datos debe tener curiosidad para explorar datos y descubrir patrones inesperados, y capacidad para mantenerse actualizado con las tendencias y tecnologías emergentes en Ciencia de Datos.

Metodologías de Análisis:

En la Ciencia de Datos, existen varias metodologías y enfoques para analizar conjuntos de datos complejos y extraer información valiosa. Estas metodologías proporcionan un marco estructurado para abordar proyectos de Ciencia de Datos, desde la comprensión del problema hasta la presentación de los hallazgos. La elección de la metodología dependerá de la naturaleza del problema, los datos disponibles y los objetivos específicos del análisis. Vamos a enfocarnos en dos de las más utilizadas en ciencia de datos (Soria, 2022).

Análisis Exploratorio de Datos (EDA)

El Análisis Exploratorio de Datos (EDA) es una fase fundamental en la Ciencia de Datos que implica explorar y comprender la estructura, patrones y relaciones presentes en los conjuntos de datos.

Objetivos del Análisis Exploratorio de Datos:
Comprensión de los Datos:

- **Identificar Patrones:** Descubrir tendencias, distribuciones y relaciones entre variables.
- **Detectar Anomalías:** Identificar valores atípicos, datos faltantes o errores en los datos.
- **Evaluar Calidad:** Determinar la calidad de los datos y su idoneidad para el análisis.

Herramientas y Técnicas Comunes:
Visualización de Datos:

- **Gráficos:** Histogramas, diagramas de dispersión, boxplots, gráficos de barras para visualizar la distribución y relación entre variables.
- **Mapas de Calor:** Representaciones visuales de la correlación entre variables.
- **Diagramas de Caja (Boxplots):** Identificación de valores atípicos y distribuciones.

Estadísticas Descriptivas:

- **Medidas Resumen:** Media, mediana, desviación estándar para resumir la distribución de los datos.
- **Correlación:** Identificación de relaciones entre variables cuantitativas.

Preprocesamiento de Datos:

- **Limpieza:** Identificación y tratamiento de valores atípicos, datos faltantes o inconsistencias en los datos.
- **Transformación:** Normalización, discretización o ajuste de escalas para preparar datos para análisis más avanzados.

Pasos en el Análisis Exploratorio de Datos:
Inspección de Datos:

- **Revisión de la Estructura:** Número de registros, variables y tipos de datos.
- **Exploración Inicial:** Visualización rápida de algunas variables para entender su distribución y relaciones.

Análisis Profundo:

- **Exploración Detallada:** Análisis más profundo de las variables clave, identificación de patrones complejos y relaciones.

- **Iteración:** Realización de análisis iterativos para descubrir más información a medida que se profundiza en los datos.

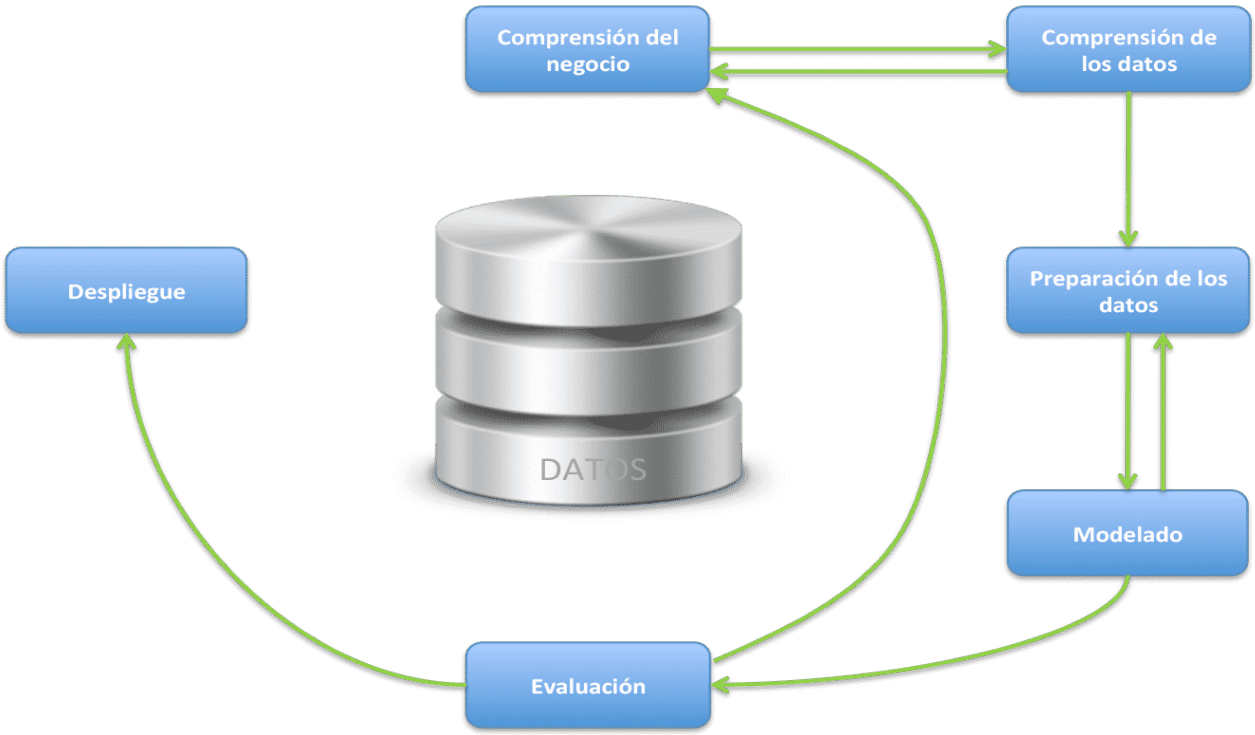
Interpretación y Comunicación:

- **Hallazgos Significativos:** Comunicación de patrones y relaciones clave descubiertos durante el análisis.
- **Próximos Pasos:** Identificación de áreas que requieren más análisis o enfoque en fases posteriores del proyecto.

El Análisis Exploratorio de Datos proporciona una base sólida para el desarrollo de modelos posteriores y la toma de decisiones informadas. Es una etapa crucial para entender la naturaleza y la calidad de los datos antes de aplicar técnicas más avanzadas de análisis y modelado.

La Metodología CRISP DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) es un marco metodológico ampliamente utilizado en proyectos de Ciencia de Datos y Minería de Datos. Consiste en seis fases interrelacionadas que guían el ciclo de vida completo de un proyecto de análisis de datos.



1. Comprensión del Negocio:

- **Objetivos del Proyecto:** Comprender los objetivos empresariales y cómo se vinculan con el análisis de datos.
- **Requisitos del Negocio:** Identificar las necesidades, los criterios de éxito y las limitaciones del proyecto.

2. Comprensión de los Datos:

- **Recopilación de Datos:** Obtener los datos necesarios para el proyecto.
- **Exploración Inicial:** Familiarizarse con los datos, entender su estructura y contenido.

- **Evaluación de la Calidad de los Datos:** Identificar problemas como valores faltantes, inconsistencias o errores.

3. Preparación de los Datos:
- **Limpieza de Datos:** Tratar con valores atípicos, datos faltantes o incoherencias.
- **Transformación de Datos:** Realizar cambios en la estructura o formato para su análisis.
- **Selección de Datos:** Seleccionar los datos relevantes para el análisis.

4. Modelado:
- **Selección de Técnica de Modelado:** Seleccionar las técnicas de modelado más apropiadas para alcanzar los objetivos del proyecto.
- **Entrenamiento del Modelo:** Desarrollar y entrenar modelos con los datos disponibles.
- **Evaluación del Modelo:** Evaluar y validar la eficacia de los modelos construidos.

5. Evaluación:
- **Evaluación del Resultado:** Valorar si los modelos alcanzan los objetivos del proyecto y cumplen con los criterios de éxito establecidos.
- **Revisión del Proceso:** Identificar posibles mejoras o pasos adicionales para mejorar el rendimiento del modelo.

6. Despliegue:
- **Implementación:** Poner en práctica los resultados del análisis de datos en el entorno operativo.
- **Seguimiento:** Monitorear el rendimiento del modelo implementado y realizar ajustes si es necesario.

Las características clave de CRISP-DM es que es adaptable a diferentes tipos de proyectos y entornos empresariales, permite iteraciones en las fases para refinar el análisis y mejorar los resultados, y las fases no son lineales y pueden influenciarse entre sí a lo largo del proyecto.

CRISP-DM proporciona una guía estructurada para gestionar proyectos de Ciencia de Datos, permitiendo a los equipos abordar cada fase de manera ordenada y eficiente para alcanzar los objetivos del proyecto de análisis de datos.

Modelado y Predicción

En la Ciencia de Datos, el modelado y la predicción son fases fundamentales que implican la construcción de modelos matemáticos y estadísticos para comprender datos pasados y predecir resultados futuros (Soria, 2022).

El objetivo del **modelado** en ciencia de datos es crear representaciones matemáticas en base a las relaciones encontradas entre variables en los datos. De esta manera, identifica estructuras subyacentes o comportamientos en los datos.

Los pasos por dar en el modelado son:

1. Definir el Problema y Objetivos: Comprender qué se quiere predecir y por qué es relevante para el negocio o el problema en cuestión, y **d**efinir métricas de rendimiento claras que se utilizarán para evaluar la precisión del modelo.

2. Recopilar y Preparar los Datos: Identificar y recopilar los datos necesarios para el modelado los datos deben ser preprocesados, tratar valores atípicos, datos faltantes, normalizar o codificar variables según sea necesario.

3. Selección del Modelo: Seleccionar el tipo de modelo adecuado según la naturaleza del problema (regresión, clasificación, etc.). Además, se debe Considerar técnicas como la regularización para evitar el sobreajuste (overfitting) del modelo.

4. División de Datos: Separar los datos en conjuntos de entrenamiento y prueba para entrenar y evaluar el modelo, respectivamente. Para ello se pueden utilizar técnicas como k-fold cross-validation para evaluar la estabilidad y generalización del modelo.

5. Entrenamiento del Modelo: Ajustar los parámetros del modelo para optimizar su rendimiento y utilizar datos de entrenamiento para que el modelo aprenda patrones y relaciones entre las variables.

6. Evaluación del Modelo: Utilizar métricas como precisión, recall, F1-score (en clasificación) o error cuadrático medio (en regresión) para evaluar la precisión del modelo. Es habitual evaluar varios modelos para seleccionar el más apropiado según las métricas establecidas.

7. Optimización y Ajuste: Refinar el modelo seleccionado mediante ajustes adicionales y validaciones. **Este proceso implica r**epetir el proceso, ajustando hiperparámetros y mejorando el modelo según sea necesario.

8. Despliegue y Monitoreo: Implementar el modelo en el entorno operativo y supervisar el rendimiento del modelo en producción y realizar ajustes según sea necesario.

A partir de aquí, con el modelo en producción comienza la fase de **predicción**, cuyo propósito es precisamente predecir resultados futuros utilizando modelos sobre datos no vistos, o dato limpio. Las predicciones son evaluadas utilizando medidas como precisión, exactitud, sensibilidad o especificidad.

Los modelos y predicciones respaldan la toma de decisiones informadas y estratégicas y ayudan a asignar recursos de manera más eficiente. Permiten ofrecer productos o servicios más personalizados según los patrones identificados.

El modelado y la predicción son aspectos esenciales en la Ciencia de Datos, ya que permiten aprovechar la información histórica para comprender el presente y predecir posibles escenarios futuros, impulsando la toma de decisiones y la innovación en diversos campos.

Estudios de Caso y Ejemplos Prácticos

Aquí tienes algunos ejemplos de casos reales en los que la Ciencia de Datos ha tenido un impacto significativo:

1. Predicción del Clima:
- **Objetivo:** Utilizar modelos predictivos para predecir patrones climáticos.
- **Aplicación:** Ayuda en la planificación agrícola, prevención de desastres naturales y gestión de recursos.

2. Análisis de Sentimientos en Redes Sociales:
- **Objetivo:** Analizar millones de comentarios en redes sociales para comprender las opiniones de los usuarios.
- **Aplicación:** Empresas usan esta información para mejorar productos, campañas de marketing y toma de decisiones estratégicas.

3. Diagnóstico Médico Asistido por Datos:
- **Objetivo:** Utilizar datos de pacientes y análisis avanzado para ayudar en el diagnóstico médico.
- **Aplicación:** Identificación temprana de enfermedades, análisis de imágenes médicas y personalización de tratamientos.

4. Recomendación de Contenido y Productos:
- **Objetivo:** Utilizar algoritmos para recomendar contenido o productos a usuarios.
- **Aplicación:** Plataformas de streaming, comercio electrónico y motores de búsqueda.

5. Prevención del Fraude Financiero:
- **Objetivo:** Identificar patrones sospechosos en transacciones financieras.
- **Aplicación:** Detectar y prevenir fraudes en tarjetas de crédito, transacciones bancarias, etc.

6. Optimización de la Cadena de Suministro:
- **Objetivo:** Utilizar análisis predictivo para mejorar la eficiencia en la gestión de inventario y logística.
- **Aplicación:** Reducción de costos, mejor planificación y entrega más rápida de productos.

7. Personalización de Experiencias en Internet:
- **Objetivo:** Utilizar datos del comportamiento del usuario para personalizar experiencias en línea.
- **Aplicación:** Ofrecer contenido relevante, publicidad personalizada y recomendaciones de productos.

Estos casos muestran cómo la Ciencia de Datos se aplica en diversas industrias para resolver problemas complejos, optimizar procesos y mejorar la toma de decisiones. Estos ejemplos ilustran la versatilidad y el impacto significativo que puede tener la aplicación de técnicas de Ciencia de Datos en el mundo real.

Referencias

Aguilar, L. J. (2016). Big Data, Análisis de grandes volúmenes de datos en organizaciones. Alfaomega Grupo Editor.

Soria, E. (2022). Inteligencia Artificial. España: RA-MA S.A. Editorial y Publicaciones.

Romero, J. A. C. (2019). Big Data. IFCT128PO. IC Editorial.