Ingeniería de Características en el Proceso de Ciencia de Datos

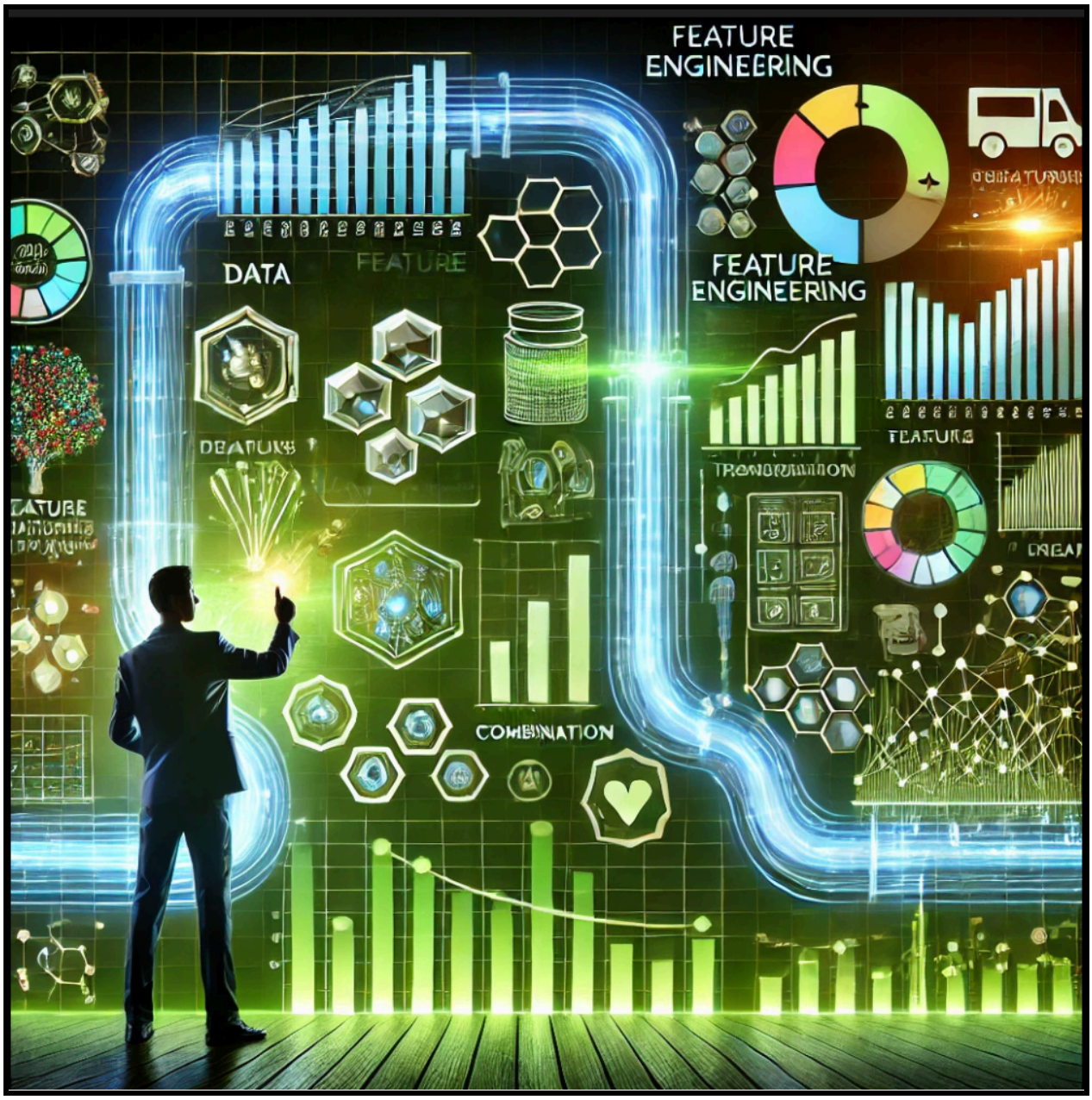
Píldora

Módulo II

Jorge Luis Mateos

Juan Domingo

FACTORÍA F5


FEATURE
ENGINEERING
DATA
FEATURE
FEATURE
ENGINEERING
COMBINATION

La Ingeniería de Características en el Aprendizaje Automático

1. Introducción a la Ingeniería de Características

La ingeniería de características es un proceso fundamental en el ámbito del aprendizaje automático. Se define como la **transformación de datos brutos en información relevante y útil que puede ser utilizada eficazmente por los modelos de aprendizaje automático**. Este proceso implica la selección, modificación y creación de variables predictivas, también conocidas como características, con el objetivo de mejorar el rendimiento y la precisión de los modelos. En esencia, la ingeniería de características se encarga de preparar los datos sin procesar en un formato que sea comprensible y procesable por los algoritmos de aprendizaje automático.

Para comprender mejor este concepto, se puede recurrir a una analogía culinaria. Imagine a un chef preparando los ingredientes antes de cocinar un plato. Los ingredientes crudos (los datos) necesitan ser limpiados, cortados y, en ocasiones, combinados de maneras específicas (las características diseñadas) para crear un plato delicioso (un modelo de alto rendimiento). De manera similar, la calidad de las características que se proporcionan a un modelo de aprendizaje automático tiene un impacto directo en su capacidad para aprender patrones y realizar predicciones precisas.

Dentro del flujo de trabajo típico de un proyecto de aprendizaje automático, la ingeniería de características se sitúa generalmente después de la recopilación y limpieza de los datos, y antes del entrenamiento del modelo. Es importante destacar que no se trata de un paso aislado, sino de un proceso iterativo que se entrelaza con la selección de datos y la evaluación del modelo. Este ciclo de ideación, creación, prueba y modificación de características continúa hasta que los datos se encuentran en un formato óptimo para que los modelos puedan generar resultados significativos.

2. La Importancia Crucial de la Ingeniería de Características en el Aprendizaje Automático

La ingeniería de características constituye un pilar fundamental en la construcción de modelos de aprendizaje automático efectivos. De hecho, el diseño y la transformación de las características de entrada pueden ser tan determinantes para el éxito de un

modelo como la elección del algoritmo en sí. Invertir tiempo y esfuerzo en la ingeniería de características puede marcar la diferencia entre un modelo de datos deficiente y uno que ofrezca predicciones precisas y conocimientos valiosos.

Un conjunto de características bien diseñado tiene un impacto significativo en el rendimiento del modelo. Se ha demostrado que una ingeniería de características eficaz puede resultar en un aumento sustancial en la precisión de los modelos, superando incluso a los algoritmos de aprendizaje más avanzados cuando se utilizan datos mal preparados. **La calidad y la relevancia de las características seleccionadas influyen directamente en la capacidad del modelo para aprender patrones subyacentes en los datos y realizar predicciones con mayor exactitud**.

Además de mejorar la precisión, la ingeniería de características contribuye a una mejor representación del problema subyacente y facilita que los algoritmos de aprendizaje automático comprendan los datos. **Al transformar los datos brutos, se pueden revelar patrones y relaciones ocultas que no serían evidentes en su forma original**. Este proceso permite que los modelos generen resultados más útiles y accionables.

Otro beneficio importante de la ingeniería de características es la posibilidad de reducir la complejidad de los modelos y acelerar el procesamiento. **Al seleccionar las características más relevantes y reducir la dimensionalidad de los datos, se pueden simplificar los modelos, lo que facilita su interpretación y mantenimiento**. Asimismo, trabajar con un espacio de características más concentrado y significativo puede llevar a tiempos de entrenamiento más rápidos y a predicciones más eficientes.

3. Técnicas Comunes de Ingeniería de Características

La ingeniería de características abarca una variedad de técnicas diseñadas para transformar y mejorar la calidad de los datos para los modelos de aprendizaje automático. A continuación, se describen algunas de las técnicas más comunes:

3.1. Manejo de Valores Faltantes

Los valores faltantes son un problema frecuente en los conjuntos de datos. Su presencia puede afectar negativamente el rendimiento de los modelos de aprendizaje automático, por lo que es crucial abordarlos de manera adecuada. Existen diversas técnicas para manejar los valores faltantes:

- **Imputación:** Esta técnica consiste en reemplazar los valores faltantes con valores estimados. Algunos métodos comunes de imputación incluyen:
    - **Imputación de la media, mediana o moda:** Para variables numéricas, los valores faltantes pueden reemplazarse con la media o la mediana de los valores existentes en esa característica. Para variables categóricas, se suele utilizar la moda, es decir, el valor más frecuente. Si bien estos métodos son sencillos, pueden introducir sesgos si la ausencia de datos no es completamente aleatoria y pueden distorsionar la varianza y las covarianzas originales.
    - **Última observación llevada hacia adelante (LOCF):** En el caso de datos de series temporales, un método común es reemplazar los valores faltantes con la última observación válida anterior. Este método asume que el valor de la variable se mantiene relativamente constante en el tiempo.
    - **Imputación basada en modelos:** Se pueden utilizar modelos de aprendizaje automático, como la regresión, para predecir los valores faltantes basándose en otras características del conjunto de datos. Este enfoque puede ser más preciso, pero también más costoso computacionalmente.
    - **Imputación múltiple:** Se crean múltiples conjuntos de datos imputados, cada uno con diferentes estimaciones para los valores faltantes, y luego se combinan los resultados para obtener una predicción final. Este método tiene en cuenta la incertidumbre introducida por los valores faltantes.
- **Eliminación:** Otra estrategia es eliminar las filas o columnas que contienen valores faltantes.
    - **Eliminación por lista (análisis de caso completo):** Se eliminan todas las filas que tengan al menos un valor faltante. Este método es simple pero puede llevar a una pérdida significativa de datos si la ausencia de valores es común.
    - **Eliminación por pares (análisis de casos disponibles):** Solo se ignoran los valores faltantes para un análisis específico, utilizando los valores presentes para las demás variables. Esto conserva más información, pero los análisis pueden no ser comparables debido a los diferentes tamaños de muestra.
    - **Eliminación de variables:** Si una variable tiene un porcentaje muy alto de valores faltantes, puede ser mejor eliminar la columna completa.
- **Creación de un indicador de valor faltante:** A veces, la ausencia de un valor en sí misma puede ser informativa. En estos casos, se puede crear una nueva

variable binaria que indique si el valor de la característica original estaba presente o faltaba.

3.2. Codificación de Variables Categóricas

Muchos algoritmos de aprendizaje automático requieren que los datos de entrada sean numéricos y no pueden procesar directamente variables categóricas. Por lo tanto, es necesario convertir estas variables en representaciones numéricas. Algunas técnicas comunes de codificación incluyen:

- **Codificación One-Hot:** Para cada categoría en una variable categórica, se crea una nueva columna binaria. Si la observación pertenece a esa categoría, la columna correspondiente tendrá un valor de 1; de lo contrario, tendrá un valor de 0. Esta técnica es adecuada para variables categóricas nominales (sin un orden inherente) pero puede aumentar significativamente la dimensionalidad del conjunto de datos si la variable tiene muchas categorías.
- **Codificación de Etiquetas (Codificación Ordinal):** Se asigna un número entero único a cada categoría en la variable. Esta técnica es apropiada para variables categóricas ordinales (con un orden inherente), pero puede introducir una relación de orden artificial en variables nominales, lo cual podría ser interpretado incorrectamente por el modelo.
- **Codificación Binaria:** Cada categoría se convierte primero en un número entero, y luego ese entero se transforma en su representación binaria. Cada dígito binario se convierte en una nueva columna. Esta técnica puede ser útil para reducir la dimensionalidad en comparación con la codificación one-hot, especialmente para variables con muchas categorías.
- **Codificación de Frecuencia:** Las categorías se reemplazan por su frecuencia relativa en el conjunto de datos. Esto puede ser útil cuando la frecuencia de una categoría es una característica predictiva importante.
- **Codificación de Destino:** Para cada categoría, se calcula la media de la variable objetivo (en problemas de regresión) o la probabilidad de la clase objetivo (en problemas de clasificación) para las observaciones que pertenecen a esa categoría, y se utiliza este valor para reemplazar la categoría. Si bien esta técnica puede capturar la relación entre la variable categórica y el objetivo, es propensa al sobreajuste y requiere una validación cuidadosa.

3.3. Escalado de Variables Numéricas

Las variables numéricas en un conjunto de datos a menudo tienen diferentes escalas y unidades de medida. Esto puede causar problemas para algunos algoritmos de aprendizaje automático que son sensibles a la magnitud de las características. El

escalado de variables numéricas asegura que todas las características contribuyan por igual al modelo. Algunas técnicas comunes de escalado incluyen:

- **Normalización (Escalado Min-Max):** Los valores de las características se escalan a un rango específico, generalmente entre 0 y 1. La fórmula es: Xnormalizado = (X - Xmin) / (Xmax - Xmin). Esta técnica es útil cuando se conocen los límites de los datos y cuando se desea mantener la forma original de la distribución. Sin embargo, es sensible a los valores atípicos.
- **Estandarización (Escalado de Puntuación Z):** Los valores de las características se transforman para que tengan una media de 0 y una desviación estándar de 1. La fórmula es: Xestandarizado = (X - μ) / σ. Esta técnica es menos sensible a los valores atípicos que la normalización y es adecuada para algoritmos que asumen una distribución gaussiana.
- **Escalado Robusto:** Similar a la estandarización, pero utiliza la mediana y el rango intercuartílico en lugar de la media y la desviación estándar. Esto hace que sea más robusto frente a la presencia de valores atípicos en los datos.
- **Transformación Logarítmica:** Se aplica una función logarítmica a los valores de la característica para reducir la asimetría en los datos. Esto puede ser útil para mejorar el rendimiento de modelos lineales.

3.4. Creación de Nuevas Características

La creación de nuevas características a partir de las existentes es una de las técnicas más poderosas de la ingeniería de características. Esto implica combinar, transformar o derivar información de las características originales para generar nuevas variables que puedan tener una mayor capacidad predictiva. Algunas formas comunes de crear nuevas características incluyen:

- **Interacciones entre características:** Se crean nuevas características combinando dos o más características existentes. Por ejemplo, se pueden multiplicar dos variables numéricas para capturar una relación interactiva entre ellas.
- **Características polinómicas:** Se generan nuevas características elevando las características numéricas existentes a diferentes potencias (por ejemplo, al cuadrado, al cubo). Esto puede ayudar a modelar relaciones no lineales.
- **Discretización (Binning):** Las características numéricas continuas se convierten en características categóricas discretas agrupando los valores en intervalos o "bins". Esto puede ayudar a simplificar la relación entre la característica y la variable objetivo y hacer que el modelo sea más robusto a los valores atípicos.
- **Extracción de características de datos complejos:**

    - **De fechas:** Se pueden extraer componentes como el día de la semana, el mes, el año, la hora del día, o si una fecha corresponde a un día festivo.
    - **De texto:** Se pueden utilizar técnicas de procesamiento del lenguaje natural para extraer características como la frecuencia de palabras clave, la presencia de n-gramas (secuencias de palabras), o métricas de sentimiento.
    - **De imágenes:** Se pueden extraer características como bordes, texturas, o utilizar modelos pre-entrenados para obtener representaciones de las imágenes.
- **Agregación:** Se pueden crear nuevas características agregando información de múltiples puntos de datos. Por ejemplo, calcular el promedio de las compras de un cliente durante un período de tiempo.

4. Ejemplos Prácticos de Ingeniería de Características

La ingeniería de características se aplica en una amplia variedad de problemas de aprendizaje automático. A continuación, se presentan algunos ejemplos prácticos:

- **Detección de Spam (Clasificación):** En este problema, el objetivo es clasificar los correos electrónicos como spam o no spam. Algunas características originales podrían ser el texto del correo electrónico, la dirección del remitente y la hora de envío. A través de la ingeniería de características, se pueden crear nuevas variables como la presencia de ciertas palabras clave (por ejemplo, "gratis", "descuento"), el número de letras mayúsculas, la longitud del correo electrónico, el dominio del remitente y la hora del día en que se envió el correo (agrupada en intervalos). Estas características diseñadas capturan patrones comunes en los correos electrónicos no deseados.
- **Predicción de Precios de Vivienda (Regresión):** El objetivo es predecir el precio de una vivienda basándose en diversas características. Las características originales podrían incluir la superficie en metros cuadrados, el número de dormitorios y baños, y la ubicación. Se pueden diseñar nuevas características como el precio por metro cuadrado, la antigüedad de la propiedad, la distancia a la escuela o al transporte público más cercano, características de interacción (por ejemplo, superficie multiplicada por el número de dormitorios) y la ubicación codificada utilizando coordenadas geográficas o la proximidad a servicios. Estas características diseñadas reflejan mejor los factores que influyen en el valor de una propiedad.

- **Análisis de Sentimiento (Clasificación):** En este caso, se busca determinar el sentimiento expresado en un texto, como una reseña de un cliente. La característica original es el texto de la reseña. Se pueden crear características diseñadas como el número de palabras positivas o negativas, la presencia de palabras específicas con carga emocional, n-gramas (secuencias de palabras) y puntuaciones de sentimiento derivadas de diccionarios de léxico. Estas características cuantifican el tono emocional del texto.
- **Detección de Fraude (Clasificación):** El objetivo es identificar transacciones fraudulentas. Las características originales podrían ser el monto de la transacción, la hora, la ubicación y la información del usuario. Se pueden diseñar características como el monto de la transacción en relación con el promedio de transacciones del usuario, la frecuencia de las transacciones en un corto período de tiempo, la distancia entre transacciones consecutivas y los patrones de gasto del usuario en el momento de la transacción. Estas características ayudan a identificar comportamientos anómalos que podrían indicar fraude.

Tipo de Problema	Características Originales	Características Diseñadas	Snippet(s)
Clasificación (Spam)	Texto del correo, dirección del remitente, hora de envío	Presencia de palabras clave, número de mayúsculas, longitud del correo, dominio del remitente, hora de envío agrupada	2
Regresión (Vivienda)	Metros cuadrados, dormitorios, baños, ubicación	Precio por metro cuadrado, antigüedad, distancia a servicios, interacciones, codificación de ubicación	29
Clasificación (Sent.)	Texto de reseñas de clientes	Conteo de palabras positivas/negativas, presencia de palabras de sentimiento, n-gramas, puntuaciones de	2

		sentimiento	
Clasificación (Fraude)	Monto, hora, ubicación, información del usuario de la transacción	Monto relativo, frecuencia, distancia entre transacciones, patrones de gasto del usuario	2

5. Desafíos Comunes en la Ingeniería de Características

A pesar de su importancia, la ingeniería de características presenta varios desafíos comunes:

- **Selección de las características correctas:** Determinar qué características son las más relevantes y contribuirán más al rendimiento del modelo puede ser un proceso complejo que a menudo requiere experiencia en el dominio y experimentación. No todas las características posibles son útiles, e incluir características irrelevantes puede incluso perjudicar el rendimiento del modelo.
- **Riesgo de sobreajuste:** Crear demasiadas características específicas, especialmente basadas en patrones observados únicamente en los datos de entrenamiento, puede llevar al sobreajuste. En este escenario, el modelo funciona muy bien con los datos de entrenamiento, pero su rendimiento se deteriora significativamente con datos nuevos o no vistos.
- **Necesidad de conocimiento del dominio:** Una ingeniería de características efectiva a menudo requiere una comprensión profunda del área problemática para identificar y crear características significativas. La colaboración con expertos en el campo puede ser invaluable para obtener información sobre qué aspectos de los datos son más propensos a ser predictivos.
- **Proceso iterativo y subjetivo:** La ingeniería de características es inherentemente un proceso iterativo que implica probar diferentes técnicas y combinaciones de características. También puede ser subjetivo, ya que diferentes científicos de datos pueden proponer conjuntos de características distintas para el mismo problema.

6. Mejores Prácticas para una Ingeniería de Características Efectiva

Para abordar los desafíos mencionados y realizar una ingeniería de características efectiva, se recomienda seguir algunas mejores prácticas:

- **Importancia del Análisis Exploratorio de Datos (EDA):** Antes de comenzar la ingeniería de características, es crucial comprender a fondo los datos a través del EDA. Esto incluye visualizar los datos, identificar patrones, comprender la distribución de las características y detectar posibles problemas como valores atípicos. El EDA proporciona la base para tomar decisiones informadas sobre qué técnicas de ingeniería aplicar.
- **Validación cruzada para evitar el sobreajuste:** Durante el entrenamiento y la evaluación del modelo, se debe utilizar la validación cruzada para evaluar qué tan bien generalizan las características diseñadas a datos no vistos. Esto ayuda a mitigar el riesgo de sobreajuste y asegura que las mejoras observadas en los datos de entrenamiento se traduzcan en un mejor rendimiento en datos nuevos.
- **Iteración y experimentación:** Es importante probar diferentes técnicas de ingeniería de características y combinaciones de características. Se debe evaluar su impacto en el rendimiento del modelo y iterar en función de los resultados. La ingeniería de características es un proceso empírico donde diferentes enfoques deben ser probados sistemáticamente.
- **Colaboración con expertos del dominio:** Involucrar a expertos en el área problemática en el proceso de ingeniería de características puede proporcionar conocimientos valiosos y guiar la creación de características más significativas y relevantes. Su conocimiento puede revelar características potencialmente útiles que podrían no ser obvias al analizar los datos por sí solos.
- **Técnicas de selección de características:** Después de diseñar nuevas características, se pueden utilizar técnicas de selección de características para identificar las más relevantes y reducir la dimensionalidad. Esto puede ayudar a prevenir el sobreajuste y mejorar la eficiencia del modelo. Existen varios métodos de selección de características, como métodos de filtro, métodos de envoltura y métodos integrados.
- **Mantenimiento de un almacén de características:** Para proyectos a gran escala, se puede considerar el uso de un almacén de características para gestionar y reutilizar las características diseñadas en diferentes modelos y pipelines. Esto mejora la coherencia y la eficiencia en la gestión de las

características.

7. Herramientas y Bibliotecas Populares para la Ingeniería de Características

En el campo del aprendizaje automático, existen varias herramientas y bibliotecas populares que facilitan el proceso de ingeniería de características, especialmente en el lenguaje de programación Python:

- **Pandas:** Proporciona estructuras de datos y herramientas de análisis de datos de alto rendimiento y fáciles de usar. Es fundamental para la manipulación y el análisis de datos, incluyendo el manejo de valores faltantes, la transformación de datos y la creación de nuevas características.
- **NumPy:** Es una biblioteca fundamental para la computación numérica en Python. Se utiliza para realizar operaciones con arrays y transformaciones matemáticas que a menudo son necesarias en la ingeniería de características.
- **Scikit-learn:** Ofrece una amplia gama de herramientas para el preprocesamiento de datos, incluyendo módulos para la imputación de valores faltantes (`impute`), la codificación de variables categóricas (`preprocessing.OneHotEncoder`, `preprocessing.LabelEncoder`) y el escalado de características numéricas (`preprocessing.StandardScaler`, `preprocessing.MinMaxScaler`). También incluye herramientas para la selección de características (`feature_selection`).
- **Feature-engine:** Es una biblioteca de Python diseñada específicamente para la ingeniería de características. Proporciona una colección de transformadores para diversas tareas, con un enfoque en una sintaxis consistente y fácil de usar.

8. Reflexión sobre la Importancia de la Ingeniería de Características

La ingeniería de características es un paso crítico en el proceso general de análisis de datos y ciencia de datos. Va más allá del simple aprendizaje automático, ya que prepara los datos para un modelado y análisis efectivos, sirviendo como puente entre los datos brutos y los conocimientos prácticos. La calidad de las características influye directamente en la fiabilidad y la perspicacia de los análisis resultantes.

Si bien la elección del modelo de aprendizaje automático adecuado es importante, unas características bien diseñadas a menudo pueden tener un impacto mayor en los resultados finales. De hecho, buenas características pueden permitir que incluso los modelos más simples funcionen de manera efectiva. La ingeniería de características y la selección de modelos deben considerarse aspectos complementarios en la construcción de sistemas predictivos eficaces.

Aunque algunos aspectos de la ingeniería de características pueden automatizarse (como se ve en las funcionalidades de AutoML), el proceso a menudo requiere creatividad, experiencia en el dominio y una comprensión profunda de los datos. Esto sugiere que, a pesar de los avances en la automatización, la ingeniería de características seguirá siendo una habilidad crucial para los científicos de datos en el futuro.

Conclusión

La ingeniería de características es un arte y una ciencia fundamental en el aprendizaje automático. Al transformar datos sin procesar en características significativas, se optimiza el rendimiento de los modelos predictivos y se facilita la obtención de conocimientos valiosos. El proceso implica una variedad de técnicas para el manejo de valores faltantes, la codificación de variables categóricas, el escalado de variables numéricas y la creación de nuevas características. Si bien presenta desafíos como la selección de las características adecuadas y el riesgo de sobreajuste, seguir las mejores prácticas, como el análisis exploratorio de datos, la validación cruzada y la colaboración con expertos del dominio, puede conducir a una ingeniería de características efectiva. El uso de herramientas y bibliotecas populares en Python simplifica la implementación de estas técnicas. En última instancia, la ingeniería de características es un componente esencial en el contexto más amplio del análisis de datos y la ciencia de datos, que permite convertir la información cruda en conocimientos accionables.