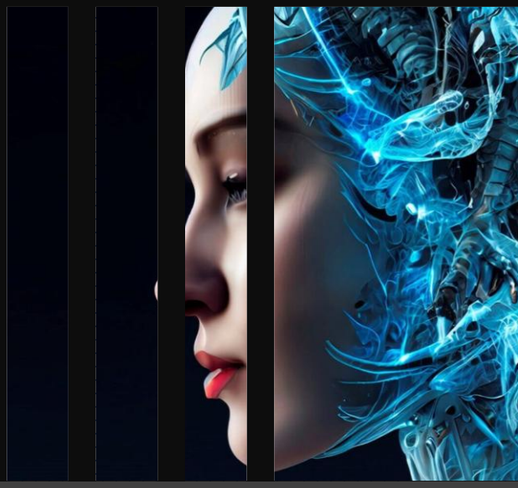
Ingeniería de selección de características

El puente esencial entre datos crudos y modelos de aprendizaje automático altamente efectivos.



Gabriel Rengifo

Objetivos de esta Sesión

En esta sesión, desglosaremos los conceptos clave y las técnicas prácticas para optimizar sus conjuntos de datos y mejorar el rendimiento de sus modelos.

Comprender la Ingeniería de Características

Definir y entender el rol crucial del *Feature Engineering* en el ciclo de vida del Machine Learning.

Explorar Técnicas de Transformación

Aprender métodos efectivos para crear y transformar variables que potencien la capacidad predictiva.

Dominar la Selección de Variables

Identificar y aplicar estrategias para elegir las características más relevantes, optimizando el rendimiento y la eficiencia del modelo.

¿Qué es una Característica (Feature)?

En el contexto del aprendizaje automático, una **característica** es una variable de entrada individual que se utiliza para predecir un resultado. Es un atributo medible de una instancia de sus datos.

> La calidad y relevancia de las características son determinantes directos del éxito y la precisión de cualquier modelo de aprendizaje automático.

Ejemplo: Para un modelo que predice la supervivencia en el Titanic, características clave serían **Edad, Sexo, Clase del Pasajero** y **Tarifa** pagada.

Ingeniería de Características (Feature Engineering)

La ingeniería de características es el arte y la ciencia de **crear, transformar o combinar** variables existentes en el conjunto de datos para mejorar significativamente el rendimiento de los modelos de aprendizaje automático.

Transformación Categórica

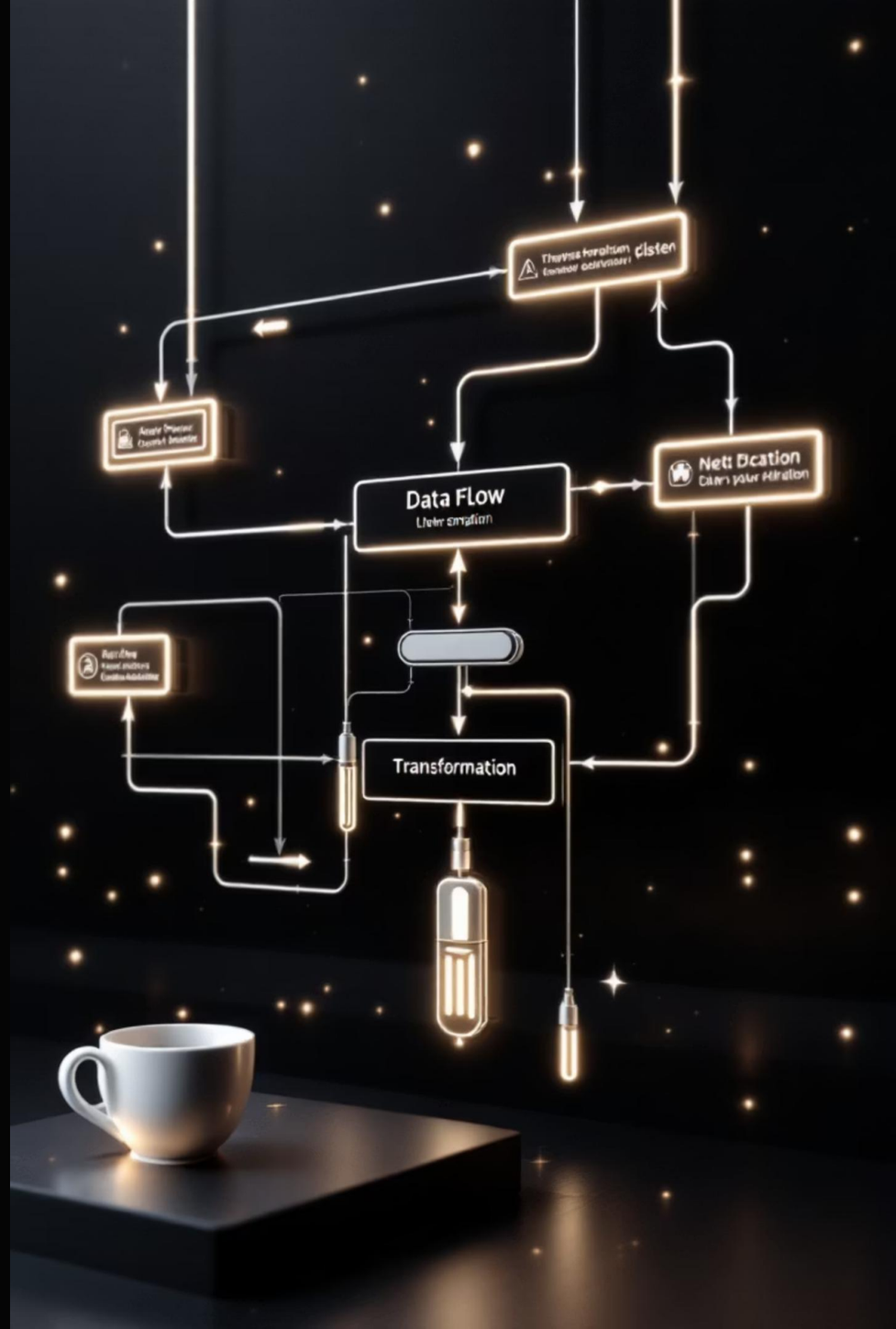
Convertir la 'Edad' numérica en categorías como 'niño', 'adulto' o 'anciano'.

Extracción Temporal

Derivar 'Día de la semana', 'Mes' o 'Estación' de una columna de 'Fecha'.

Vectorización de Texto

Aplicar TF-IDF (Frecuencia de Término - Frecuencia Inversa de Documento) o conteo de palabras para representar texto numéricamente.



Beneficios Clave de un Buen Feature Engineering

Precisión Mejorada

Los modelos pueden aprender patrones más complejos y relevantes.

Reducción de Ruido

Elimina información redundante o irrelevante que puede confundir al modelo.

Relaciones Ocultas

Permite descubrir y explotar conexiones no obvias en los datos.

Mayor Interpretabilidad

Características bien definidas facilitan la comprensión del comportamiento del modelo.

Técnicas Comunes de Ingeniería de Características

Codificación de Variables Categóricas

Convertir etiquetas textuales en formatos numéricos, como **One-Hot Encoding** o **Label Encoding**, esenciales para muchos algoritmos.

Tratamiento de Outliers

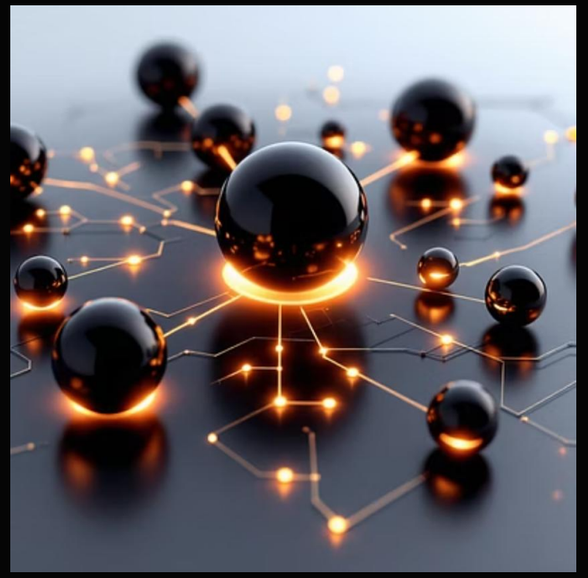
Manejar valores atípicos mediante técnicas como **Winsorización** (reemplazo por valores extremos aceptables) o **Transformaciones Logarítmicas**.

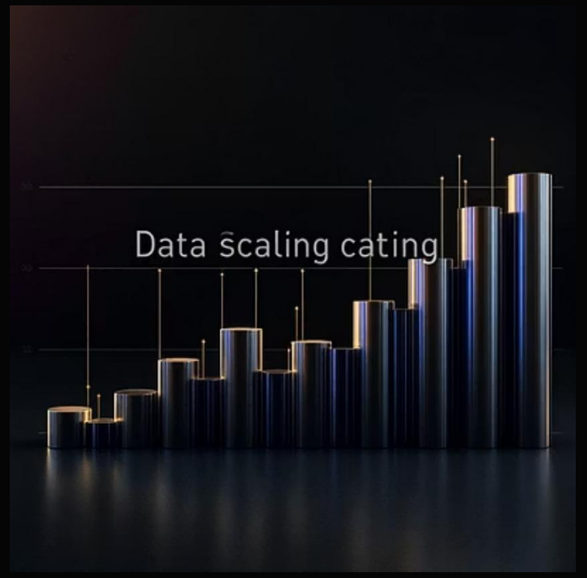
Escalamiento y Normalización

Ajustar rangos de valores con **Min-Max Scaling** o **Z-score Standardization** para asegurar que ninguna característica domine desproporcionadamente.

Creación de Nuevas Variables

Generar características más informativas a partir de las existentes: **variables polinomiales, combinaciones, ratios** o **interacciones**.

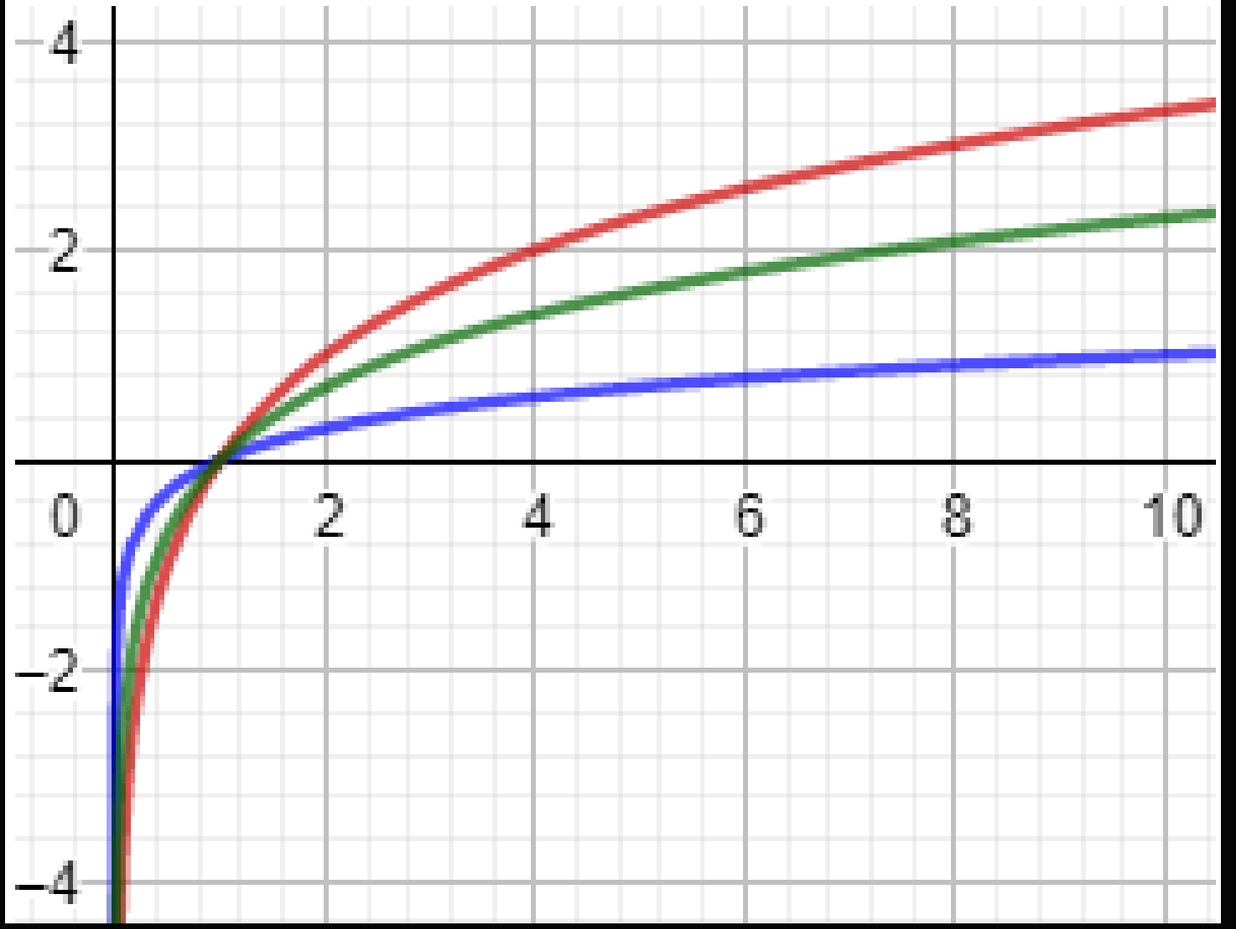




$$f(x) = \log_2(x)$$

$$f(x) = \ln(x)$$

$$f(x) = \log_{10}(x)$$



One-Hot Encoding

Favorite Color	Height (m)	Loves Troll 2
Blue	1.77	Yes
Red	1.32	No
Green	1.81	Yes
Blue	1.56	No
Green	1.64	Yes
Green	1.61	No
Blue	1.73	No

Blue	Red	Green	Height (m)	Loves Troll 2
1	0	0	1.77	1
0	1	0	1.32	0
0	0	1	1.81	1
1	0	0	1.56	0
0	0	1	1.64	1
0	0	1	1.61	0
1	0	0	1.73	0

Label Encoding

Favorite Color	Height (m)	Loves Troll 2
0	1.77	1
1	1.32	0
2	1.81	1
0	1.56	0
2	1.64	1
2	1.61	0
0	1.73	0

Target Encoding

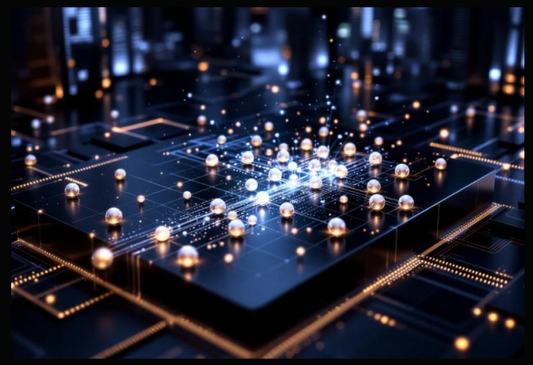
Favorite Color	Height (m)	Loves Troll 2
0.33	1.77	1
0	1.32	0
0.67	1.81	1
0.33	1.56	0
0.67	1.64	1
0.67	1.61	0
0.33	1.73	0

Selección de Características (Feature Selection)

La selección de características es el proceso de reducir el conjunto de variables de entrada a aquellas que son más relevantes para la predicción, optimizando así el rendimiento y la eficiencia del modelo.

Métodos de Filtro

Evalúan la relación entre cada característica y la variable objetivo de forma independiente, utilizando estadísticas como **correlación, Chi-cuadrado** o **ANOVA**. Son rápidos y robustos.

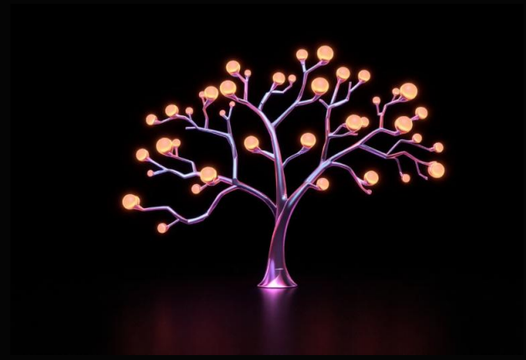


Métodos de Envoltura (Wrapper)

Evalúan subconjuntos de características utilizando un algoritmo de aprendizaje automático específico. Un ejemplo es **Recursive Feature Elimination (RFE)**.

Métodos Embebidos (Embedded)

La selección se realiza como parte del proceso de entrenamiento del modelo. Ejemplos incluyen la **importancia de características en modelos de árboles** o la regularización **Lasso/Ridge**.



Selección de Características (Feature Selection)

Método	Ejemplo	Pros	Contras
Filtro	Correlación, Chi^2, ANOVA	Rápidos, fáciles	No capturan interacciones
Envoltura	RFE con LogisticRegression	Precisión alta, considera interacciones	Lento, costoso
Embebidos	Importancia en RandomForest, Lasso	Selección automática durante entrenamiento	Dependiente del modelo

Ejemplo Práctico: Datos del Titanic

Variables Originales:

`Age` (Edad), `Sex` (Sexo), `Pclass` (Clase del Pasajero), `Fare` (Tarifa).

Estas variables proporcionan una base, pero su poder predictivo puede ser limitado por sí solas.

Nuevas Variables Derivadas:

- `Child`: Una variable binaria (0/1) indicando si la `Age` es menor de 12 años.
- `FamilySize`: Calculada como la suma de hermanos/cónyuges (`SibSp`) + padres/hijos (`Parch`) + 1 (por el propio pasajero).
- `IsAlone`: Una variable binaria que indica si `FamilySize` es igual a 1 (viaja solo).

Estas nuevas variables capturan relaciones que los modelos podrían no inferir directamente de las características originales. La selección final se haría manteniendo las más predictivas según correlación y la evaluación de modelos base.

Mejores Prácticas en Ingeniería y Selección

Documentación Exhaustiva

Registre cada transformación, su propósito y cómo fue aplicada. Esto es crucial para la reproducibilidad y el mantenimiento.

Evitar Fugas de Información (Data Leakage)

Asegúrese de que las características creadas no contengan información de la variable objetivo o del conjunto de prueba. Procese los datos de entrenamiento y prueba por separado.

Validación Rigurosa

Siempre valide el rendimiento de su modelo con el conjunto de prueba o mediante validación cruzada para asegurar que las mejoras sean generalizables y no producto de un sobreajuste.

Priorizar la Simplicidad

Un conjunto de características más pequeño y simple a menudo conduce a modelos más robustos, más rápidos y más fáciles de interpretar. Menos variables a menudo es más.

Conclusiones Clave

Arte y Ciencia

La ingeniería de características es más un arte intuitivo que una ciencia exacta; requiere creatividad y conocimiento del dominio para desbloquear el potencial de los datos.

El Poder del Feature Set

Un conjunto de características bien curado y relevante puede superar el rendimiento de un modelo intrínsecamente complejo, demostrando que la calidad de la entrada importa más que la complejidad del algoritmo.

Modelos Optimizados

Una selección de características adecuada resulta en modelos más rápidos, precisos, eficientes y, fundamentalmente, más robustos y confiables en entornos de producción.