Compresión de las bases del aprendizaje supervisado

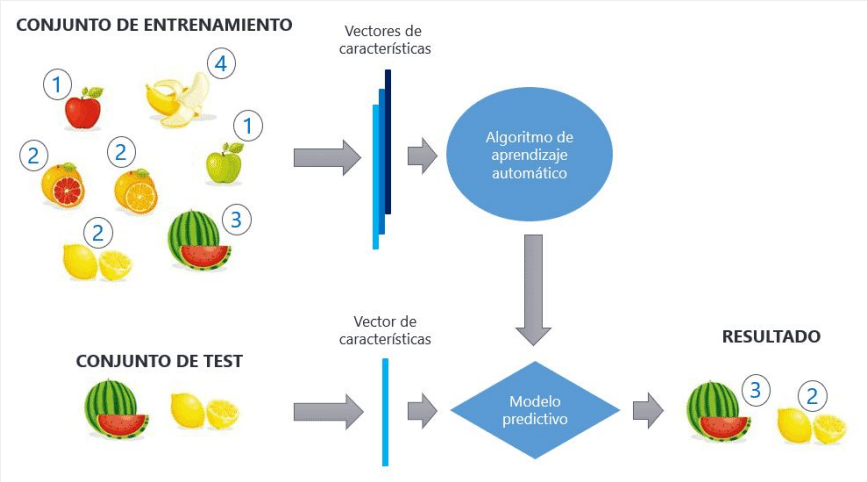
Marzo, 2025

Rafael Martín R.

Introducción al Aprendizaje Supervisado

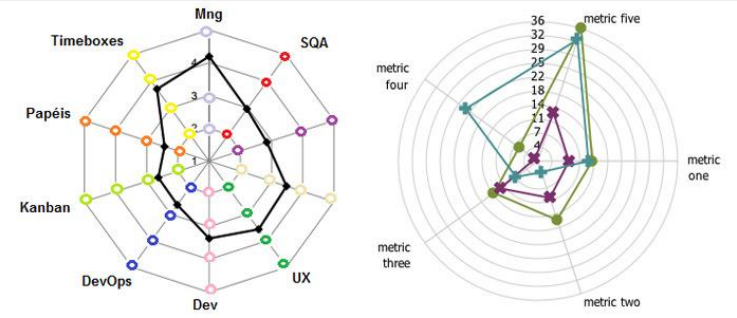
Descubriendo el poder del aprendizaje automático guiado

El aprendizaje supervisado es una rama fundamental del machine learning donde el modelo aprende a partir de datos etiquetados. Este proceso implica entrenar algoritmos con ejemplos donde conocemos las respuestas correctas, permitiendo al modelo identificar patrones y realizar predicciones precisas en nuevos datos. Es como enseñar a un estudiante mostrándole ejemplos resueltos para que luego pueda resolver problemas similares por sí mismo.



Rafael Martín R.

El Problema de Optimización



Formulación del Problema

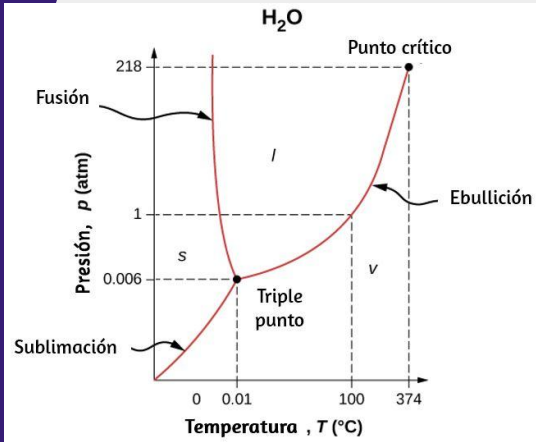
El aprendizaje supervisado busca encontrar una función que mapee correctamente las variables de entrada X a las variables objetivo Y, minimizando el error de predicción a través de ejemplos etiquetados.

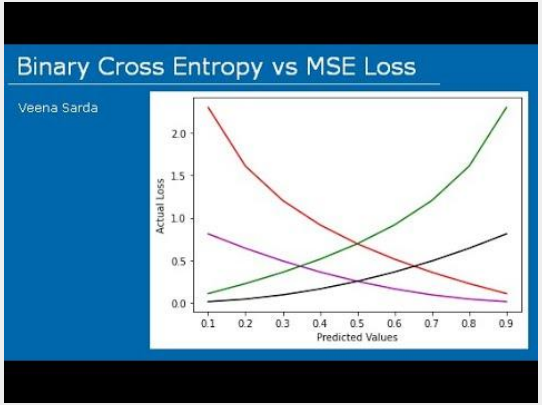
Variables y Predicción

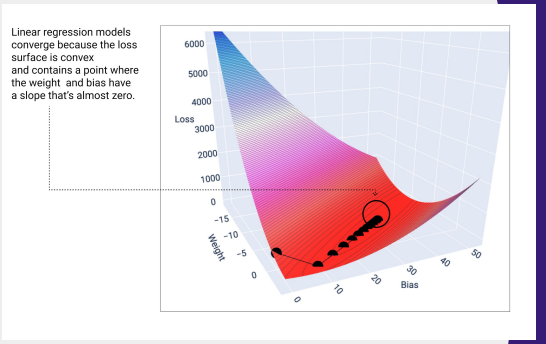
Los datos etiquetados son fundamentales para entrenar el modelo, donde cada ejemplo contiene características de entrada (features) y su correspondiente valor objetivo, permitiendo al algoritmo aprender patrones y realizar predicciones precisas.

Rafael Martín R.

Función de Pérdida







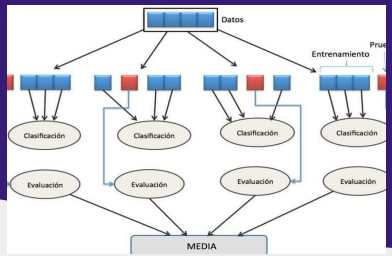
¿Qué es una función de pérdida?

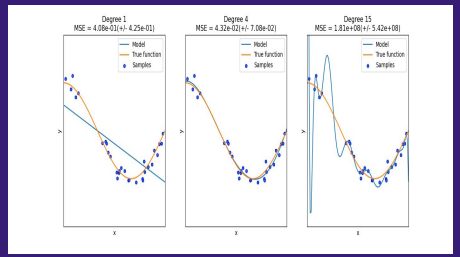
Tipos de funciones de pérdida

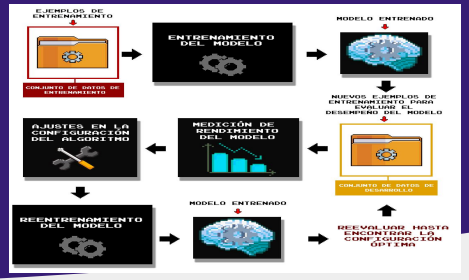
Optimización y minimización

Rafael Martín R.

Conjuntos de Datos en el Aprendizaje Supervisado







Conjunto de Entrenamiento (Training Set)

- Datos principales para entrenar el modelo y ajustar parámetros internos
- Representa típicamente el 60-80% del conjunto total de datos
- Base fundamental para que el modelo aprenda patrones y relaciones

Conjunto de Validación (Validation Set)

- Evalúa el rendimiento durante el entrenamiento y ajusta hiperparámetros
- Ayuda a detectar problemas de overfitting o underfitting temprano
- Permite optimizar la arquitectura y configuración del modelo

Conjunto de Prueba (Test Set)

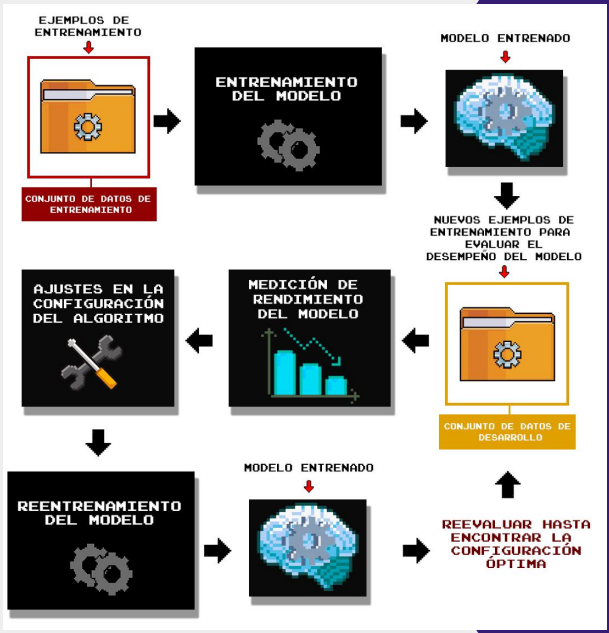
- Evalúa el rendimiento final del modelo con datos nunca vistos
- Proporciona una estimación imparcial del rendimiento en producción
- Debe mantenerse intacto hasta la evaluación final del modelo

Rafael Martín R.

Conjunto de Entrenamiento

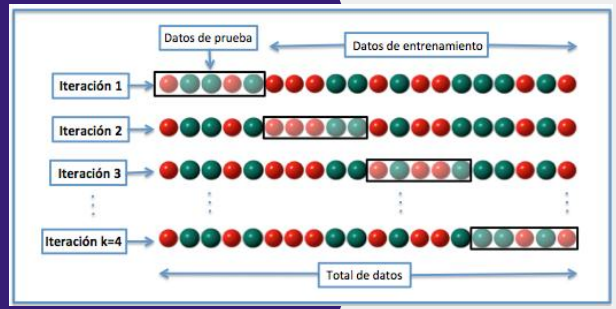
Fundamentos del conjunto de datos para entrenar modelos de ML

- El conjunto de entrenamiento es la base principal para que el modelo aprenda patrones y relaciones entre características
- La calidad y representatividad de los datos determina directamente el rendimiento final del modelo entrenado
- El preprocesamiento incluye normalización, codificación de variables categóricas y manejo de valores faltantes
- Es fundamental realizar limpieza de datos para eliminar outliers y corregir inconsistencias en los datos
- El balanceo de clases asegura que el modelo no desarrolle sesgos hacia categorías sobrerrepresentadas
- La augmentación de datos puede ayudar a incrementar el tamaño del conjunto cuando los datos son limitados
- La validación del conjunto de entrenamiento garantiza que los datos sean adecuados para el problema específico



Rafael Martín R.

Conjunto de Validación



El conjunto de validación es una herramienta crucial en el aprendizaje supervisado que nos permite evaluar el rendimiento del modelo durante el entrenamiento. A través de técnicas como la validación cruzada, podemos obtener una estimación más robusta del desempeño del modelo y ajustar sus hiperparámetros de manera óptima. Este conjunto actúa como un termómetro que nos ayuda a detectar problemas como el sobreajuste o subajuste tempranamente, permitiéndonos realizar ajustes antes de la evaluación final con el conjunto de prueba.

Rafael Martín R.

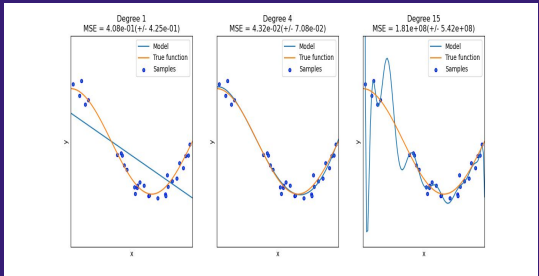
Conjunto de Prueba

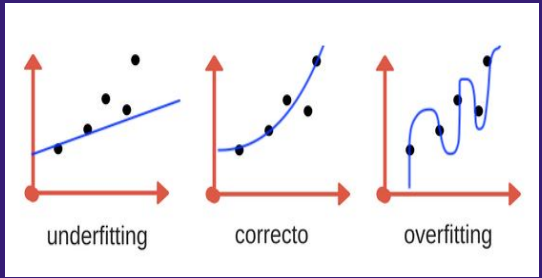
Evaluación definitiva del modelo para medir su capacidad de generalización real

- El conjunto de prueba es un conjunto de datos independiente y nunca antes visto por el modelo
- Permite evaluar de manera imparcial el rendimiento final del modelo en datos nuevos
- Ayuda a determinar la capacidad de generalización real del modelo a situaciones desconocidas
- Se utiliza una única vez al final del proceso de desarrollo del modelo
- Es fundamental mantener este conjunto completamente aislado durante todo el entrenamiento
- Las métricas obtenidas en prueba son indicadores realistas del rendimiento en producción
- Se recomienda usar entre 20-30% de los datos totales para el conjunto de prueba

Rafael Martín R.

Curva de Aprendizaje





Curva de Aprendizaje Ideal

Diagnóstico de Problemas

Rafael Martín R.

Análisis de Errores

Comprendiendo y Mejorando a través del Análisis de Errores

El análisis de errores es fundamental para mejorar el rendimiento de los modelos de aprendizaje supervisado. Los errores más comunes incluyen falsos positivos y falsos negativos, que se visualizan mediante la matriz de confusión. Esta herramienta nos permite identificar patrones específicos de error, como la confusión sistemática entre clases particulares. Al interpretar estos resultados, podemos implementar estrategias específicas como el rebalanceo de clases, la recolección de más datos para casos problemáticos, o el ajuste de hiperparámetros para mejorar el rendimiento del modelo.

Rafael Martín R.

Métricas de Evaluación - Clasificación

Accuracy

Proporción total de predicciones correctas. Ideal para datasets balanceados donde todos los errores tienen igual importancia.

Precision y Recall

Precision mide la exactitud de predicciones positivas, Recall mide la capacidad de detectar todos los casos positivos. Cruciales en detección de fraude o diagnóstico médico.

F1-Score

Media armónica entre Precision y Recall. Métrica equilibrada para datasets desbalanceados donde necesitamos balance entre exactitud y cobertura.

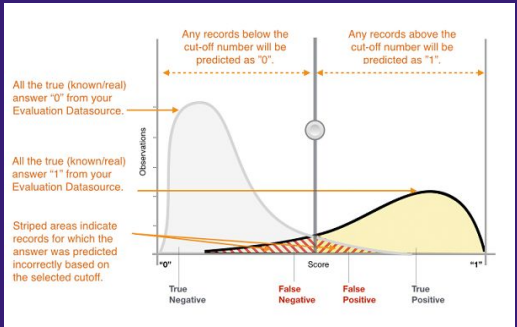
Rafael Martín R.

Métricas de Evaluación - Regresión

Error Medio Absoluto (MAE)	Error Cuadrático Medio (MSE)	Raíz del Error Cuadrático Medio (RMSE)	Coeficiente de determinación (R^2)
Promedio de diferencias absolutas entre predicciones y valores reales. Útil cuando los valores atípicos no son críticos.	Promedio de errores al cuadrado. Penaliza más fuertemente errores grandes, sensible a valores atípicos.	Raíz cuadrada del MSE. Proporciona error en las mismas unidades que la variable objetivo.	Indica qué porcentaje de la varianza es explicado por el modelo. Valores entre 0 y 1.

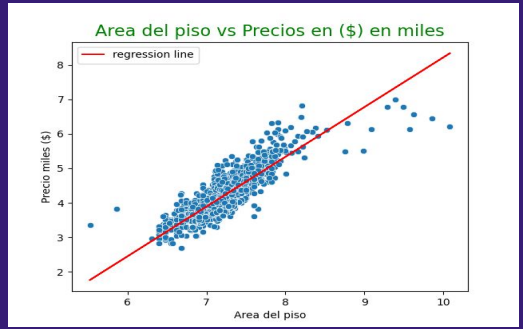
Rafael Martín R.

Casos Prácticos



Clasificación Binaria: Detección de Fraude

- Modelo logístico para detectar transacciones fraudulentas bancarias
- Accuracy del 92% y recall del 85% en datos de prueba
- Matriz de confusión muestra baja tasa de falsos positivos



Regresión: Predicción de Precios Inmobiliarios

- Modelo de regresión lineal múltiple con variables normalizadas
- RMSE de 25,000€ y R^2 de 0.85 en validación
- Residuales muestran distribución normal sin sesgos significativos

Rafael Martín R.

Buenas Prácticas en Aprendizaje Supervisado

Preparación de Datos

Limpieza exhaustiva de datos, manejo de valores faltantes, normalización de variables y codificación adecuada de características categóricas para garantizar calidad en el entrenamiento.

Selección y Validación

Elección de métricas apropiadas según el problema y dominio específico. Implementación de validación cruzada y pruebas rigurosas para asegurar la robustez del modelo.

Documentación y Reproducibilidad

Documentación detallada del proceso, incluyendo decisiones de diseño, parámetros utilizados y resultados obtenidos. Mantenimiento de código versionado y entorno reproducible.

Rafael Martín R.

Resumen y Conclusiones

- El aprendizaje supervisado requiere datos etiquetados y una función de pérdida bien definida para el entrenamiento efectivo
- La división adecuada de datos en conjuntos de entrenamiento, validación y prueba es fundamental para evaluar el rendimiento
- Las curvas de aprendizaje son herramientas esenciales para diagnosticar problemas de underfitting y overfitting
- La selección de métricas debe alinearse con los objetivos del negocio y el tipo de problema
- El análisis de errores proporciona insights valiosos para mejorar el rendimiento del modelo
- La validación cruzada y el ajuste de hiperparámetros son cruciales para optimizar el modelo
- La documentación y reproducibilidad garantizan la implementación exitosa en producción

Rafael Martín R.

Consideraciones

Consideraciones Clave

- Alinear métricas con objetivos del negocio y contexto
- Analizar errores y patrones para mejorar el modelo
- Implementar validación cruzada y ajuste de hiperparámetros
- Balancear complejidad del modelo con tamaño de datos

Desafíos a Evitar

- Depender de una única métrica de evaluación
- Ignorar señales de overfitting o underfitting
- Optimizar sin considerar el contexto del problema
- Descuidar la calidad y preparación de datos