UNIVERSIDAD NACIONAL DE LA MATANZA

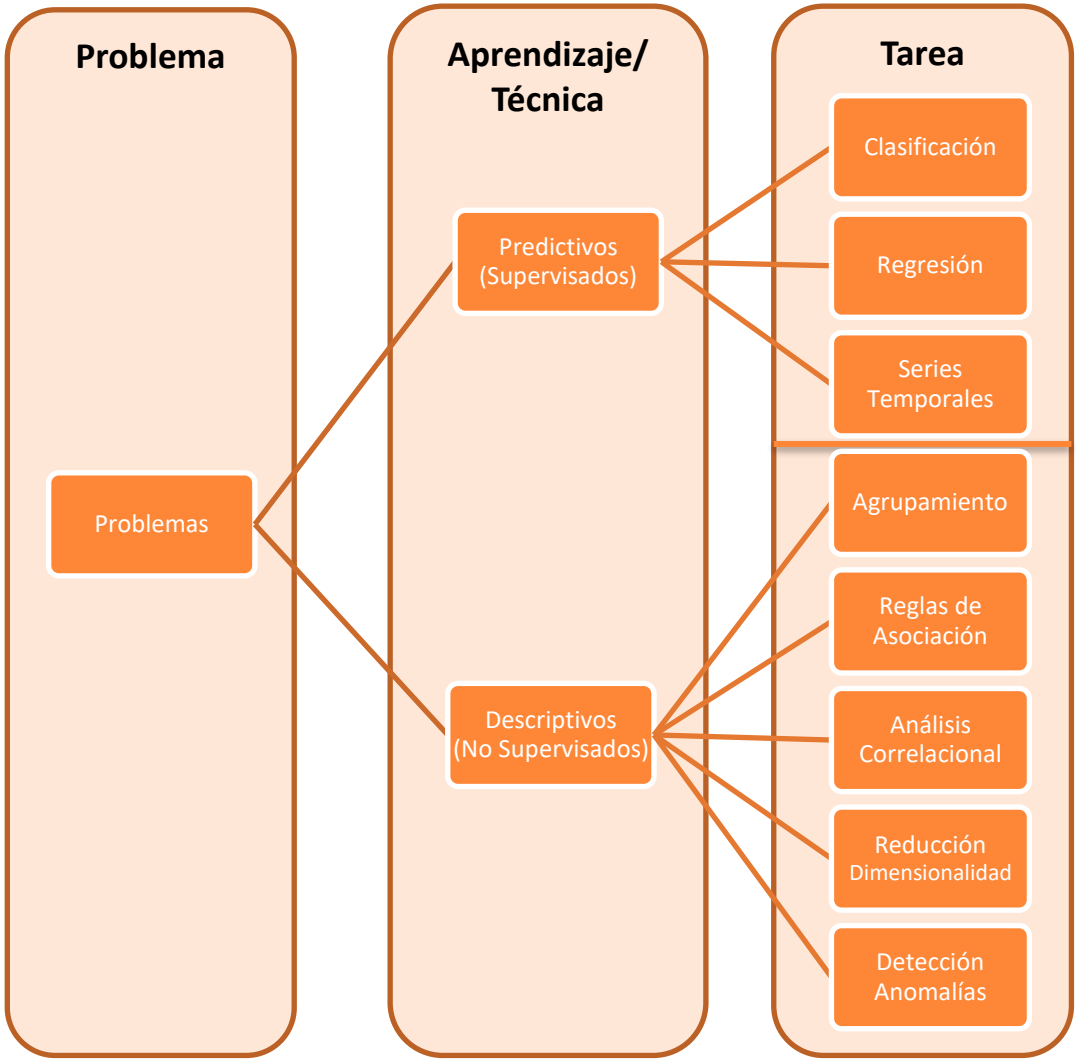
INTELIGENCIA DE NEGOCIOS

Tecnologías Inteligentes para Explotación de Información

Docente: ING. LORENA R. MATTEO

Fecha última actualización.: 18/6/2025

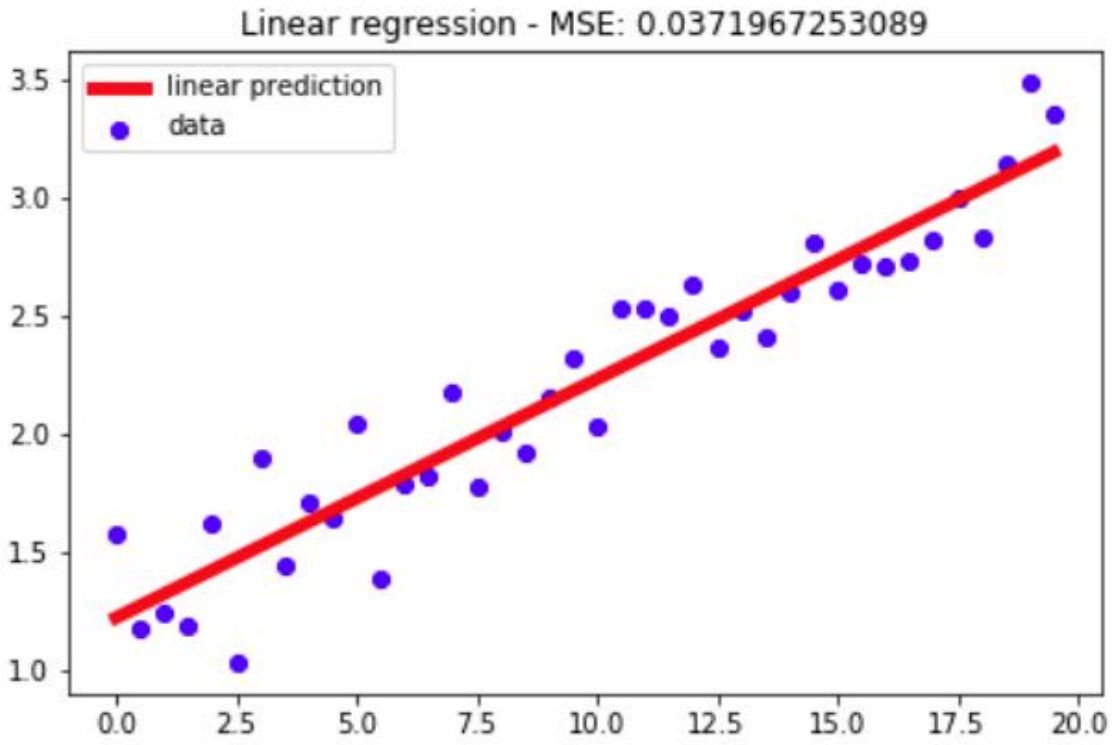
Técnicas de Minería de Datos



Las técnicas de minería de datos son herramientas que facilitan el descubrimiento de conocimiento.

¿Como aprende un modelo de Minería de Datos?

- Encontrando los valores de los hiperparámetros del modelo que minimicen el error.



m = 0.1014
b = 1.2258

La clave del éxito

¿Qué hace que un modelo sea bueno?

Un buen modelo:

- Generaliza bien con datos nuevos.
- Se evalúa con datos no usados en el entrenamiento.
- Se compara con otros modelos mediante métricas como Precisión, Recall, etc.

La clave del éxito

Se debe evaluar la capacidad de generalización del modelo.

- El modelo construido debe ser “generalizable”, debe aprender bien con muchos tipos de datos nuevos.
- Debe evaluarse con datos distintos a los usados en el entrenamiento.

¿Cómo saber si un modelo es bueno o no?

Hay que enfocarse en la capacidad predictiva y sencillez del modelo, más que en su rapidez para clasificar, construir modelos y escalar...

- Utilizando métricas de desempeño (performance metrics)
  Las métricas se calculan contrastando los valores predichos versus los valores reales de la variable objetivo.
- Comparándolo con el desempeño de otros modelos posibles, diseñados a través de experimentos.
- Probando con datos de prueba y otros dataset (generalizable, por ej. bajo error de generalización)

El problema del sobreaprendizaje (Overfitting)

¿A qué se llama sobreajuste o sobreaprendizaje?

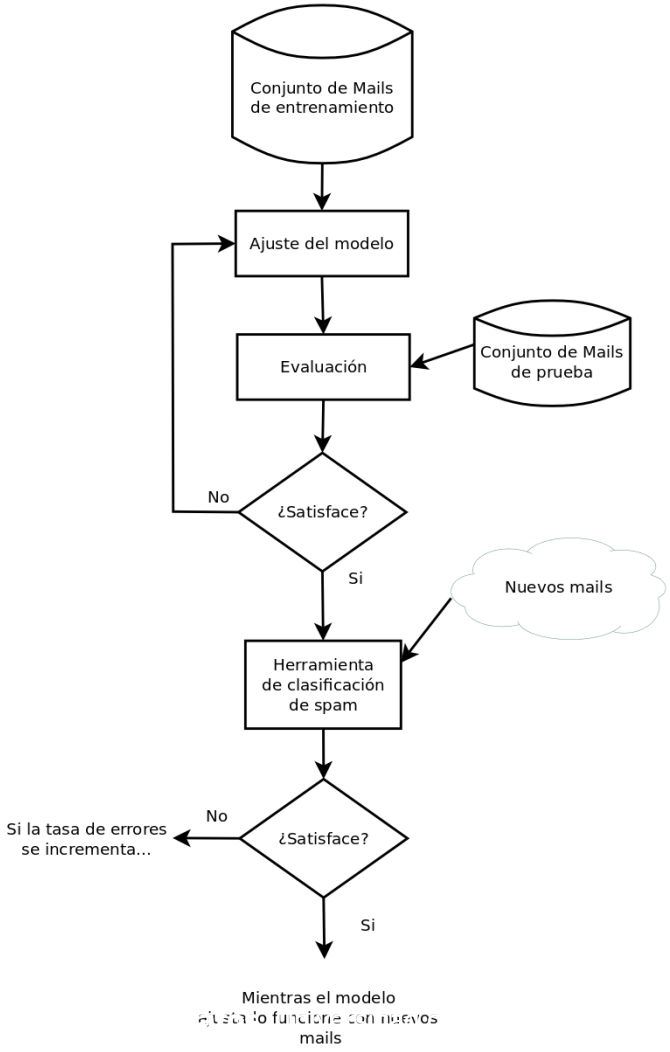
- Cuando un modelo puede aprender demasiado bien los datos de entrenamiento, pero fallar con datos nuevos.

  Por ejemplo: Cuando un estudiante memoriza respuestas sin entender el problema.

  Es decir,
  - Un modelo muy complejo que memoriza los datos de entrenamiento.
  - Tiene Bajo Error en Entrenamiento, Alto en Prueba

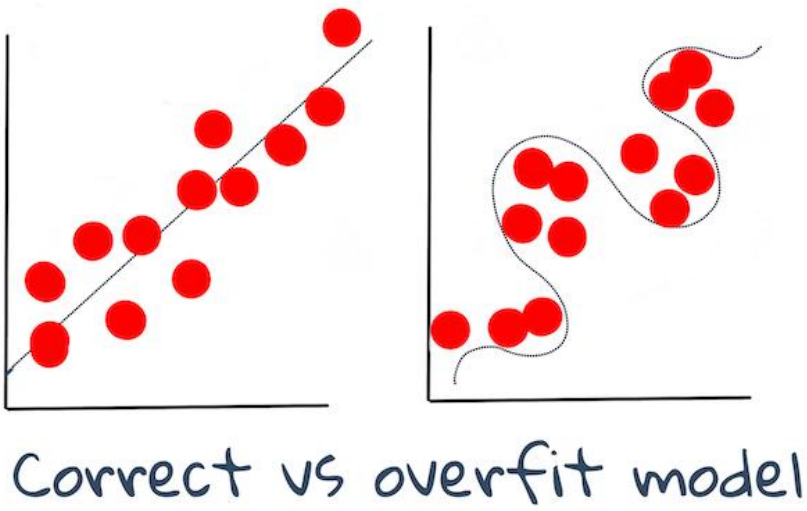
El problema del sobreaprendizaje (Overfitting)

- El propósito de crear un modelo o clasificador no es clasificar el conjunto de entrenamiento, sino para clasificar los datos cuya clase no sabemos.
- Queremos que los datos sean clasificados correctamente, pero a menudo no tenemos forma de saber si el modelo lo hace. Si la naturaleza de los datos cambia con el tiempo.
  - *Ejemplo: detectar correos electrónicos no deseados.*



El problema del sobreaprendizaje (Overfitting)

- Cuanto mayor sea su complejidad, los modelos de clasificación tienden a ajustarse más al conjunto de entrenamiento utilizado en su construcción (sobreaprendizaje), lo que los hace menos útiles para clasificar nuevos datos.
- En consecuencia, el conjunto de prueba debe ser siempre independiente del conjunto de entrenamiento.



El error de clasificación en el conjunto de entrenamiento NO es un buen estimador de la precisión del clasificador.

Generalización - Análisis de errores

¿Con qué Tipos de Errores nos encontramos?

- **Bias (Sesgo del Modelo):** modelo demasiado simple.
- **Varianza:** modelo demasiado complejo.

¿Qué pasa si sumamos datos?

- Se reduce la varianza de modelos complejos
- Se mejora la generalización
- Se evita o reduce el sobreajuste

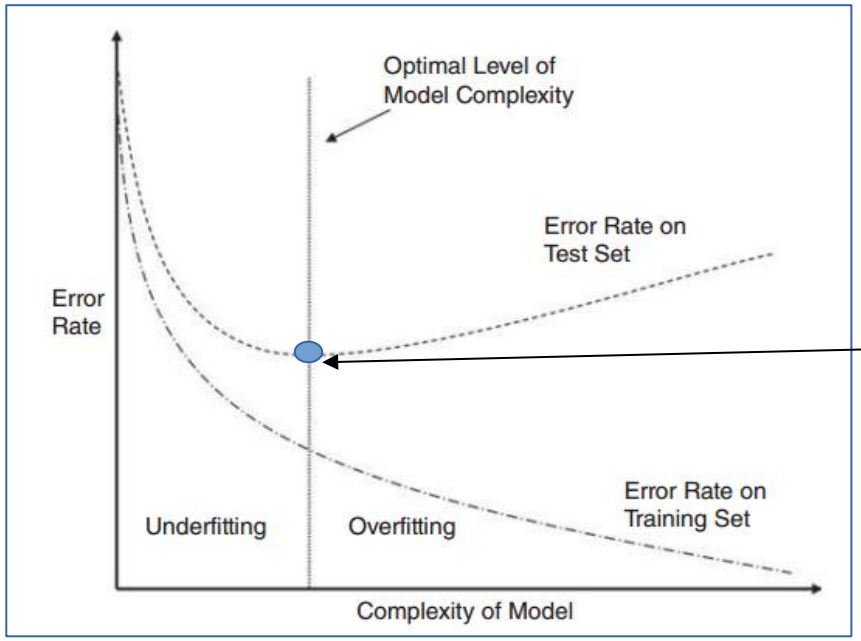
"El mejor modelo logra un equilibrio entre Sesgo y Varianza."

Generalización - Análisis de errores

Train Error	Test Error	¿Qué hacer?
LOW OVER-FITTING	HIGH	• Necesidad de un modelo más sencillo (se está empleando un modelo más complejo que lo necesario). • Se necesitan más datos (más muestras de datos).
HIGH UNDER-FITTING	HIGH	• Necesidad de un modelo más complejo. • Se necesitan más muestras de datos • Difícil aprender f(x, z) sólo con x. Obtener también z. • Agregar funciones adicionales/ hiperparámetros.
HIGH	LOW	• Inusual: podría significar que los datos de prueba son demasiado similares a los de entrenamiento. • Se necesitan más datos de prueba.
LOW	LOW	• **¡¡Lo has logrado!! ¡¡Felicitaciones!!**

El problema del sobreaprendizaje (Overfitting)

- Cuando la exactitud del modelo no es tan alta en el conjunto de prueba (evaluación) como lo es en el de entrenamiento, a menudo se debe a que el modelo sobreajusta el conjunto de entrenamiento.



Conforme el modelo se vuelve más complejo pierde capacidad de generalizar.

Mínimo error en test

Diagnóstico de generalización - Casos comunes

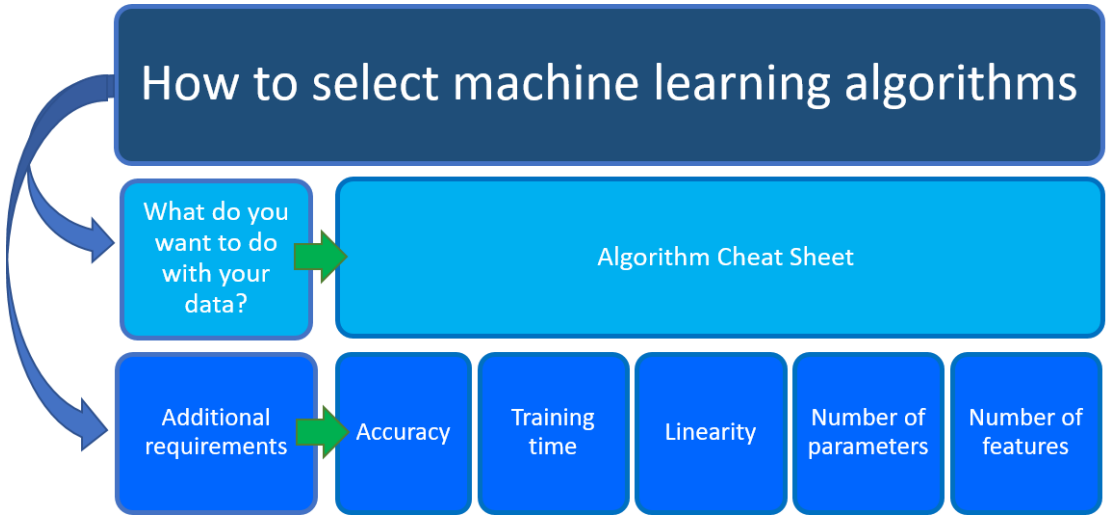
Resumen:

- **Train Error Bajo + Test Error Alto** ➜ Modelo muy complejo.
- **Ambos Errores Altos** ➜ Modelo pobre o falta de datos.
- **Ambos Bajos** ➜ Modelo ideal.

¿Qué Algoritmo de Aprendizaje Automático usar?

La pregunta es muy habitual cuando uno comienza a trabajar en Minería de Datos. Si bien no se responde de manera directa y no tiene una única respuesta, el algoritmo a seleccionar depende principalmente de dos aspectos diferentes:

- ¿Qué desea hacer con los datos? Específicamente, ¿cuál es la pregunta de negocios a responder aprendiendo de los datos disponibles?
- ¿Cuáles son los requisitos de su caso de estudio? Específicamente, ¿cuál es la precisión, el tiempo de entrenamiento, la linealidad, la cantidad de parámetros y la cantidad de características que admite su solución?



Fuente: How to select a machine learning algorithm - Azure Machine Learning | Microsoft Learn

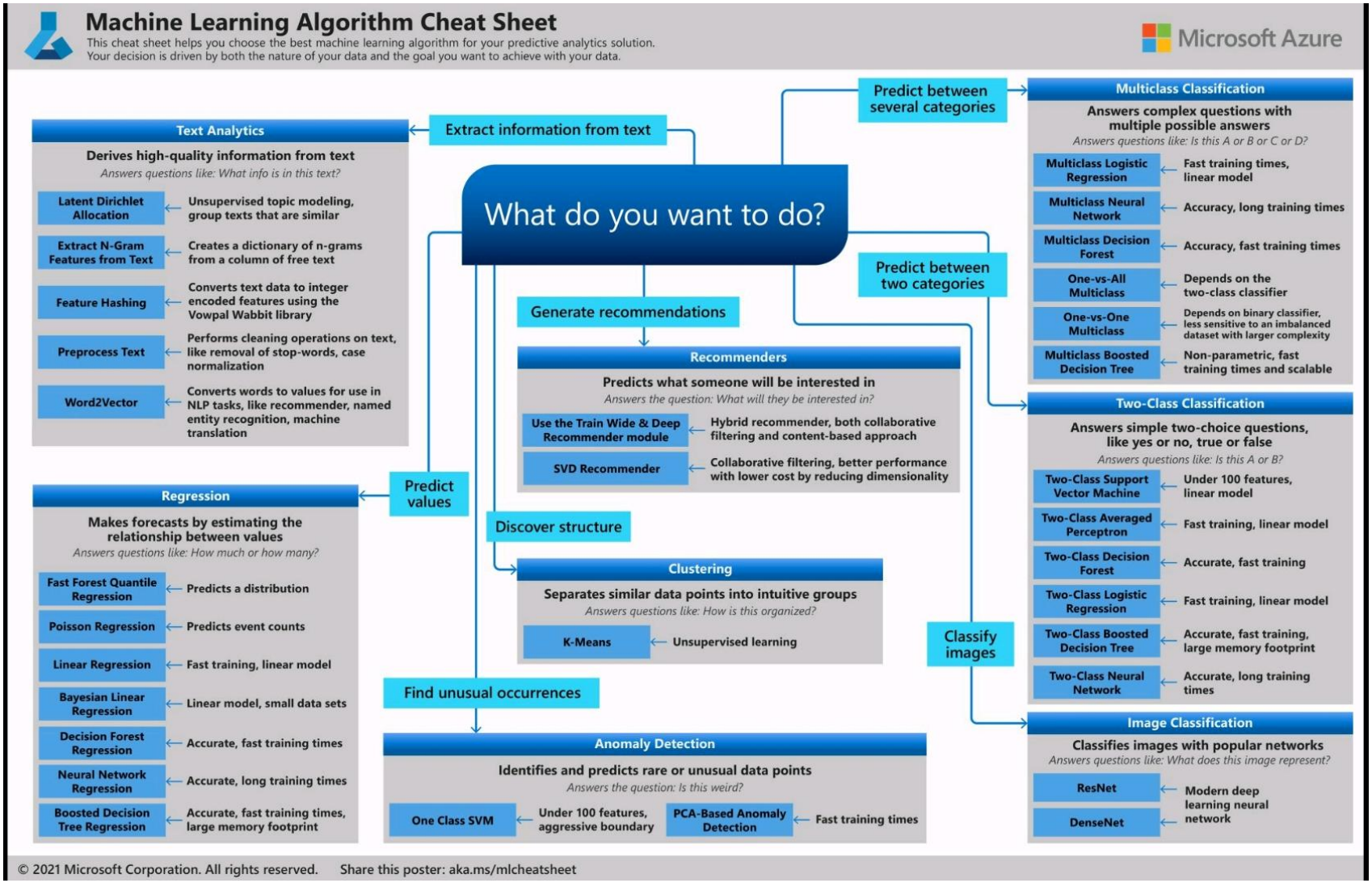
¿Qué algoritmo de aprendizaje automático debo usar?

- Algunos algoritmos de aprendizaje hacen suposiciones particulares sobre la estructura de los datos o los resultados deseados.
- Si puede encontrar uno que se adapte a sus necesidades, puede brindarle resultados más útiles, predicciones más precisas o tiempos de entrenamiento más rápidos.
- La siguiente tabla resume algunas de las características más importantes de los algoritmos de las familias de clasificación, regresión y agrupamiento:

Algorithm	Accuracy	Training time	Linearity	Parameters	Notes
Classification family					
Two-Class logistic regression	Good	Fast	Yes	4	
Two-class decision forest	Excellent	Moderate	No	5	Shows slower scoring times. Suggest not working with One-vs-All Multiclass, because of slower scoring times caused by tread locking in accumulating tree predictions
Two-class boosted decision tree	Excellent	Moderate	No	6	Large memory footprint
Two-class neural network	Good	Moderate	No	8	
Two-class averaged perceptron	Good	Moderate	Yes	4	
Two-class support vector machine	Good	Fast	Yes	5	Good for large feature sets
Multiclass logistic regression	Good	Fast	Yes	4	
Multiclass decision forest	Excellent	Moderate	No	5	Shows slower scoring times
Multiclass boosted decision tree	Excellent	Moderate	No	6	Tends to improve accuracy with some small risk of less coverage
Multiclass neural network	Good	Moderate	No	8	
One-vs-all multiclass	-	-	-	-	See properties of the two-class method selected
Regression family					
Linear regression	Good	Fast	Yes	4	
Decision forest regression	Excellent	Moderate	No	5	
Boosted decision tree regression	Excellent	Moderate	No	6	Large memory footprint
Neural network regression	Good	Moderate	No	8	
Clustering family					
K-means clustering	Excellent	Moderate	Yes	8	A clustering algorithm

Fuente: How to select a machine learning algorithm - Azure Machine Learning | Microsoft Learn

¿Qué algoritmo de aprendizaje automático debo usar?



Fuente: How to select a machine learning algorithm - Azure Machine Learning | Microsoft Learn

¿Qué Algoritmo de Aprendizaje Automático usar?

Depende de:

- Tipo de problema
- Recursos (tiempo, memoria)
- Precisión deseada

Algoritmo	Precisión	Tiempo	Linealidad
Regresión logística	Buena	Rápido	Sí
Random Forest	Excelente	Medio	No
Red neuronal	Buena	Medio	No

Referencias y Material Adicional

Referencias:

- Jose Martinez Heras | European Space Agency (ESA)
- Curso Inteligencia de Negocios, Universidad de Granada, España
- Curso DM Datos, Prof. F.Bravo/ H. Sarmiento / B. Poblete, Universidad de Chile
- Curso Minería Datos, Universidad de Luján, Argentina
- Microsoft/ML-For-Beginners
- Curso Advanced Machine Learning | Educación IT , Argentina

Material Adicional:

- Sesgo algorítmico: https://www.youtube.com/watch?v=nkOPXgpWS9w
  https://www.youtube.com/watch?v=onJs6DXczu0