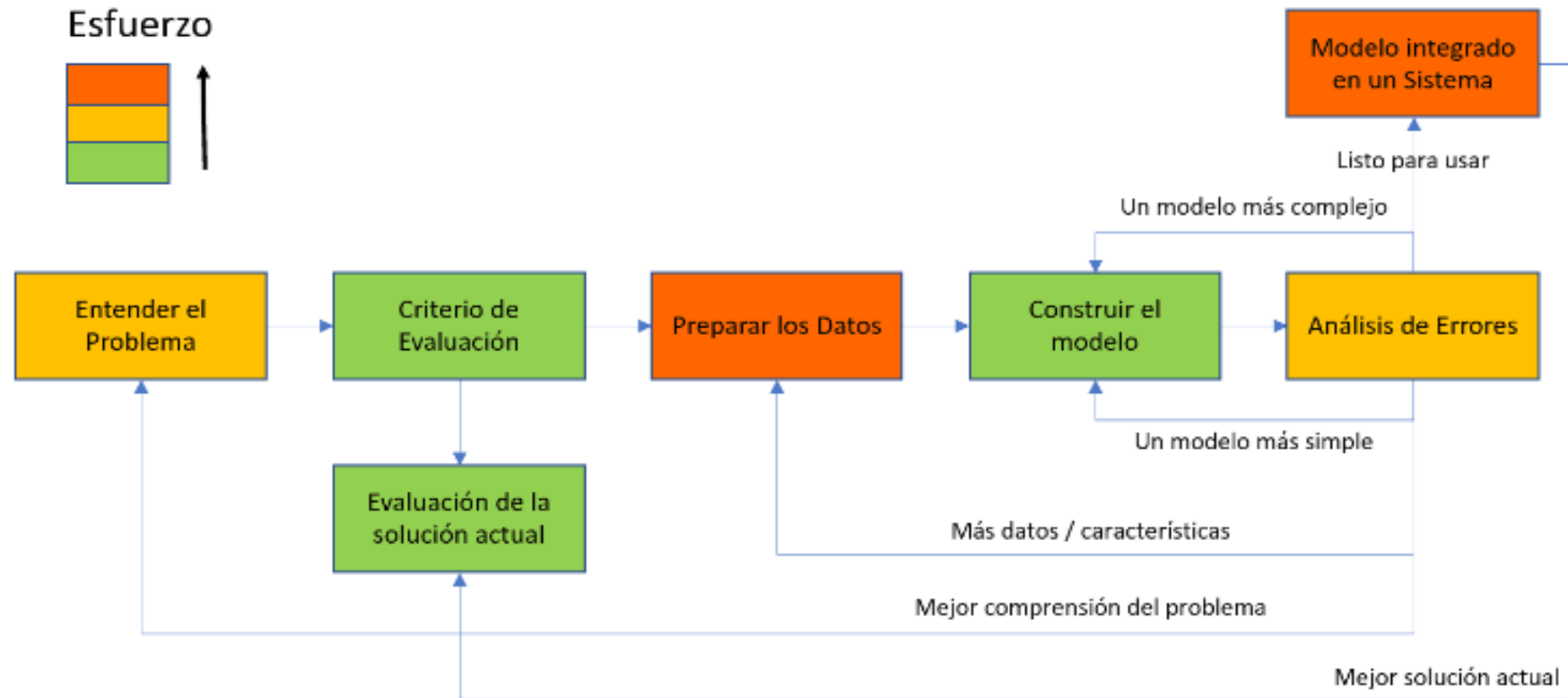


IMPLEMENTACIÓN DE MODELOS PREDICTIVOS MEDIANTE ALGORITMOS DE MACHINE LEARNING



M.I. Rodrigo Dominguez García

FASES DEL PROCESO DE MACHINE LEARNING.



Referencia: <https://www.iartificial.net/fases-del-proceso-de-machine-learning/>

COMPRENSIÓN Y DEFINICIÓN DEL ALCANCE DEL PROBLEMA.

Para poder abordar mejor nuestro problema debemos entender el alcance de las soluciones de machine Learning para tener expectativas realistas sobre los resultados que buscamos obtener.

Debemos definir cuales serán nuestras **variables de entrada o explicativas** para nuestro modelo, así como nuestra **variable de interés o de respuesta**, en este punto es importante entender como se representa nuestros objetos de estudio para los algoritmos de machine Learning (definición del **Vector característica**)

Al implementar un modelo de predicción debemos definir si el problema se va abordar como **clasificación o regresión**.

DEFINIR CRITERIOS DE EVALUACIÓN.

Al implementar un modelo de predicción debemos definir si el problema se va abordar como **clasificación** o **regresión**.

Debemos definir una **linea base** que sirva como **punto de referencia** para evaluar los resultados, de nuestros modelos de Machine Learning.

Algunas consideraciones claves que debemos tener en cuenta son:

- Evaluar Métricas de Rendimiento(**RMSE,MSE,Accuracy, precission,...**).
- Evaluación de la capacidad del aprendizaje del modelo (**Learnig Curves**).
- Comparación del rendimiento del modelo con respecto al **Baseline**.

PREPARACIÓN DE LOS DATOS.

Importancia de los datos en Machine Learning:

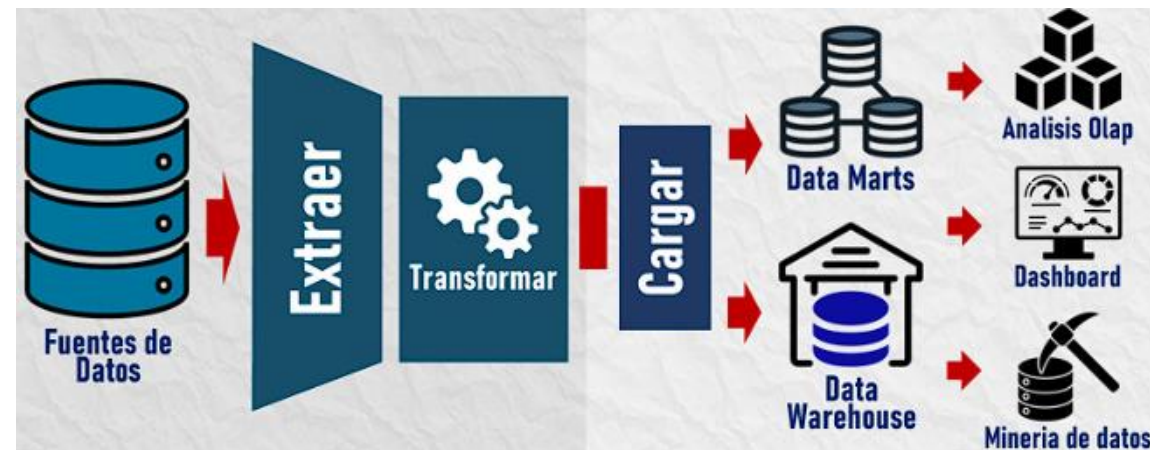
"Basura entra, basura sale" (Garbage In, Garbage Out - GIGO).

Fuentes de datos:

Bases de datos estructuradas, APIs, Sensores IoT, imágenes, texto.

Preprocesamiento:

Limpieza de datos (manejo de valores nulos, duplicados), Eliminación de outliers, Normalización y estandarización, Manejo de datos categóricos, análisis exploratorio.





CONSTRUCCIÓN DEL MODELO: REPRESENTACIÓN DE LOS DATOS EN ML.

Cada pieza de información que se incluye en la representación de nuestro problema se conoce como **características**.

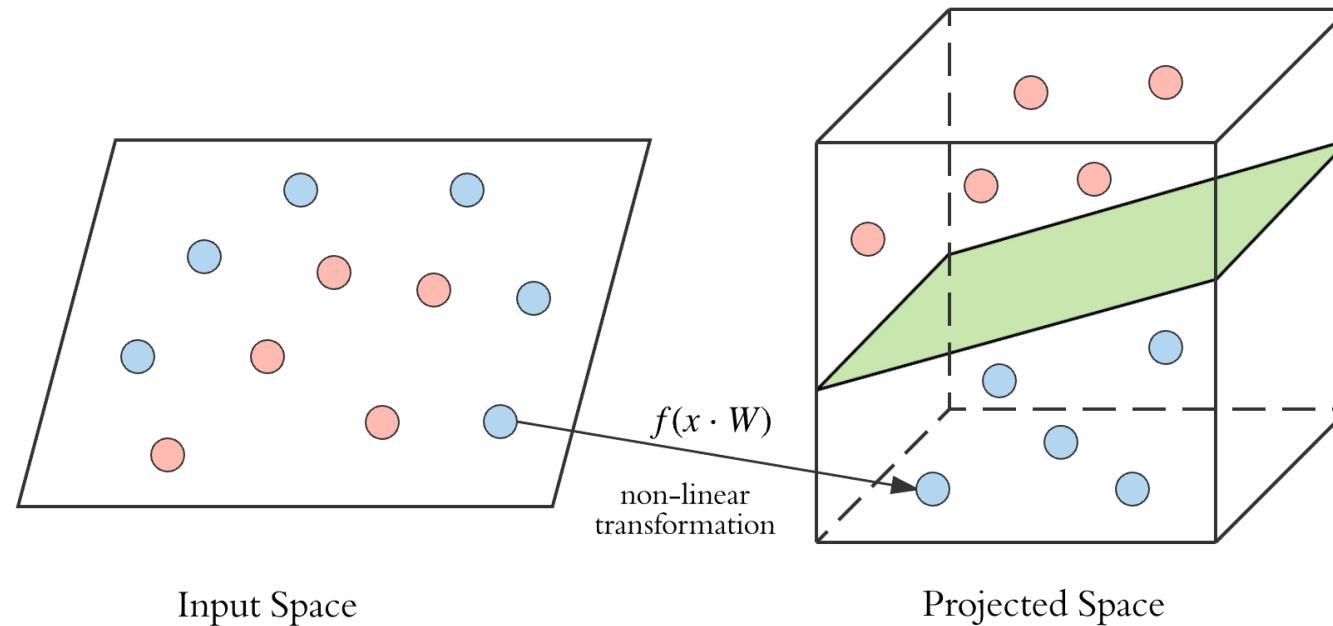
Estas características pueden ser números, vectores, matrices o tensores.



	=	<table><tr><td>.713</td><td>6.8</td><td>6.3</td></tr><tr><td>1.01</td><td>.847</td><td>1.19</td></tr><tr><td>1.36</td><td>.077</td><td>.919</td></tr></table>	.713	6.8	6.3	1.01	.847	1.19	1.36	.077	.919
.713	6.8	6.3									
1.01	.847	1.19									
1.36	.077	.919									
	=	<table><tr><td>.618</td><td>.198</td><td>3.4</td></tr><tr><td>4.05</td><td>.457</td><td>3.46</td></tr><tr><td>4.2</td><td>.734</td><td>.431</td></tr></table>	.618	.198	3.4	4.05	.457	3.46	4.2	.734	.431
.618	.198	3.4									
4.05	.457	3.46									
4.2	.734	.431									

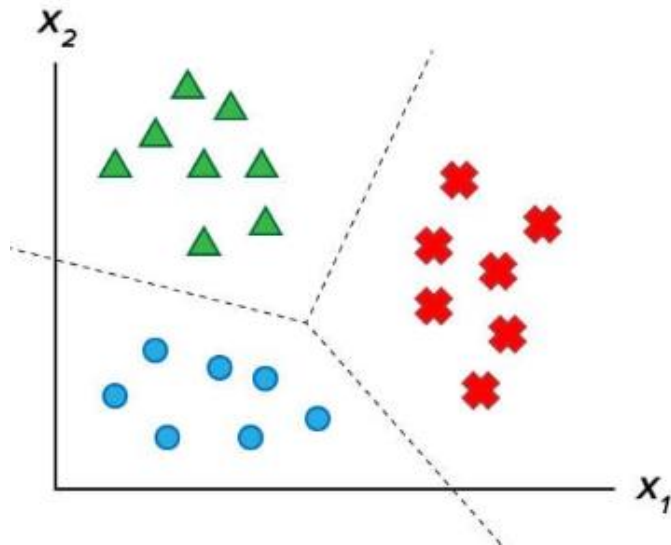
CONSTRUCCIÓN DEL MODELO: REPRESENTACIÓN DE LOS DATOS EN ML.

Se busca **representar cada objeto** o individuo de nuestros datos **mediante un vector de características**, en el cual cada característica representa una dimensión en un espacio n-dimensional, dentro de este espacio los algoritmos de ML van a buscar crear planos para separar nuestros datos.

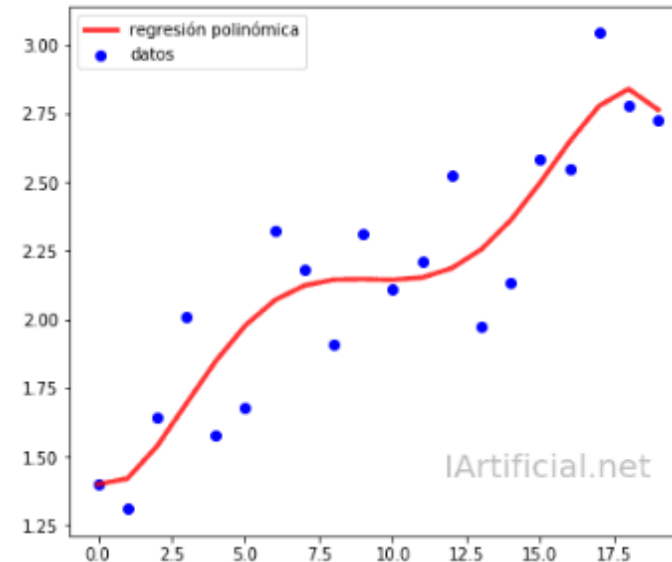


CONSTRUCCIÓN DEL MODELO: APRENDIZAJE SUPERVISADO.

Se refiere a un tipo de modelos de ML que se entrenan con un conjunto de ejemplos en los que los resultados de salida son conocidos. Hay dos aplicaciones principales de aprendizaje supervisado:



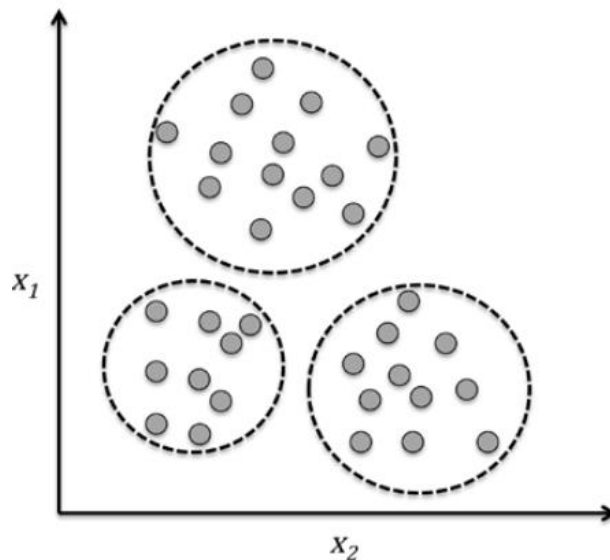
Clasificación



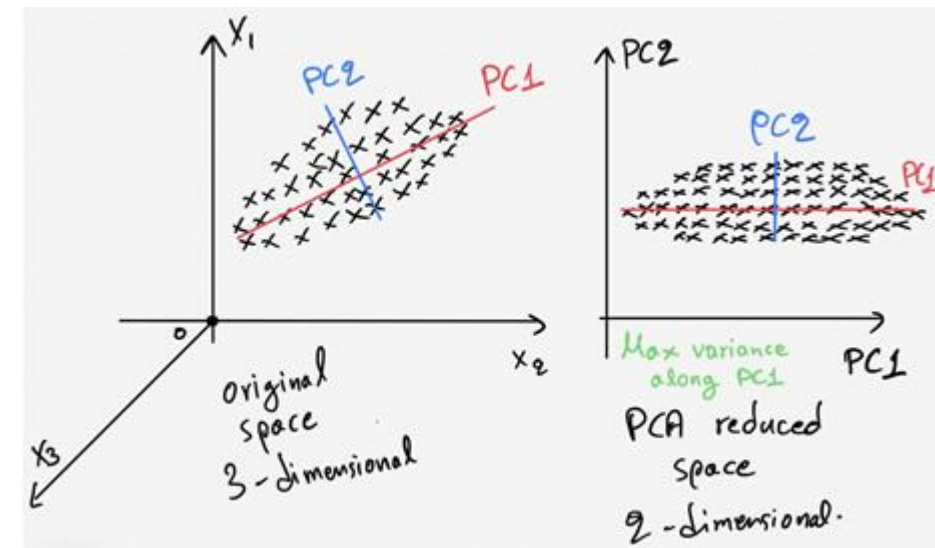
Regresión

CONSTRUCCIÓN DEL MODELO: APRENDIZAJE NO SUPERVISADO.

Se basa datos sin etiquetar cuya estructura es desconocida. El objetivo será la extracción de información significativa, sin la referencia de variables de salida conocidas, y mediante la exploración de la estructura de dichos datos sin etiquetar. Hay dos categorías principales:



Agrupamiento



Reducción de dimensionalidad

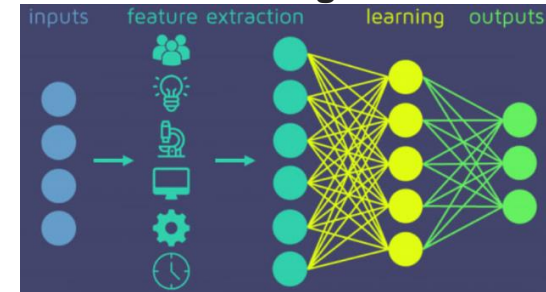
SELECCIÓN DEL MODELO DE MACHINE LEARNING.

Después de definir nuestro **vector característica** y el tipo de problema a resolver, el siguiente reto es la selección del algoritmo de Machine Learning para la creación de nuestro modelo que nos permita **minimizar el error en las predicciones** en los datos de entrenamiento y mejorar la **generalización** con los datos de validación, pero el problema es navegar entre las múltiples opciones, y el objetivo principal

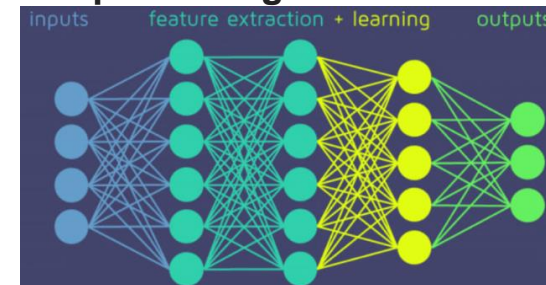
Modelos:

- Lineal Regression
- Polinomial Regresión
- Lasso, Ridge y ElasticNet Regression
- Logistic Regression
- RandomForest
- GradientBosting
- Support Vector Machine
- K-Nearest Neighbors
- Artificial neural network

Machine Learning



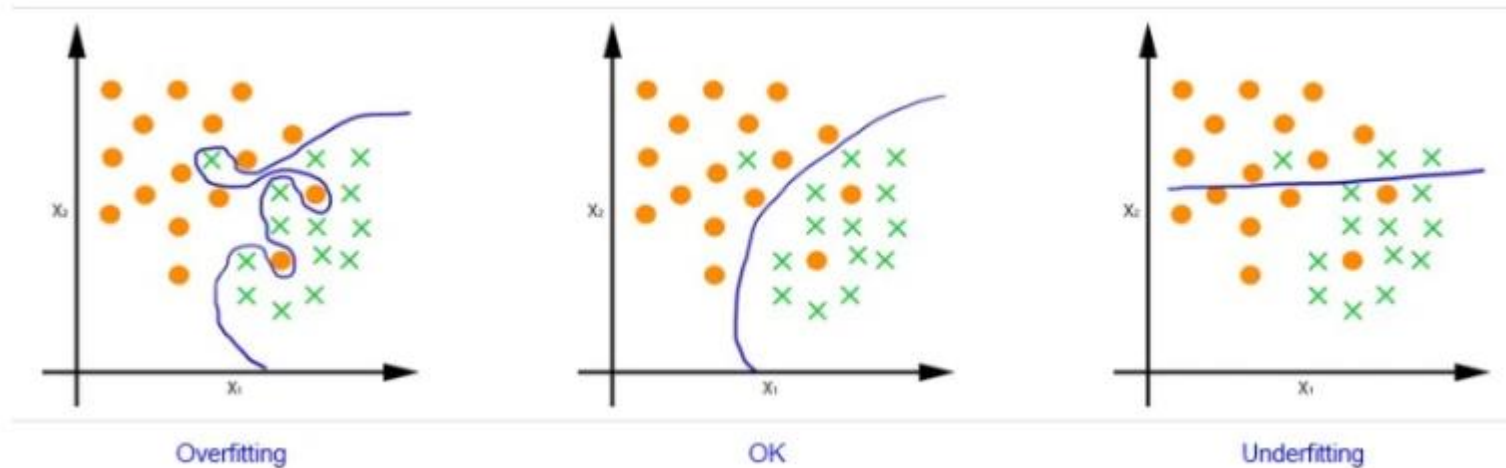
Deep Learning



ANÁLISIS DE LOS RESULTADOS DEL MODELO: GENERALIZACIÓN, OVERFITTING Y UNDERFITTING EN ML.

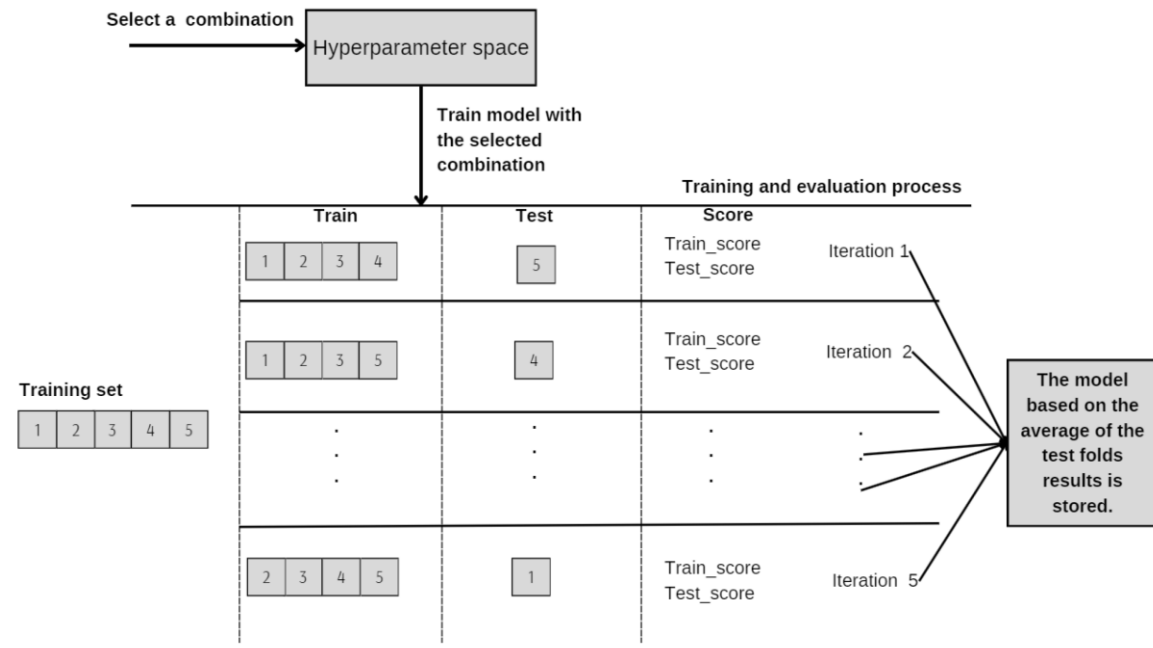
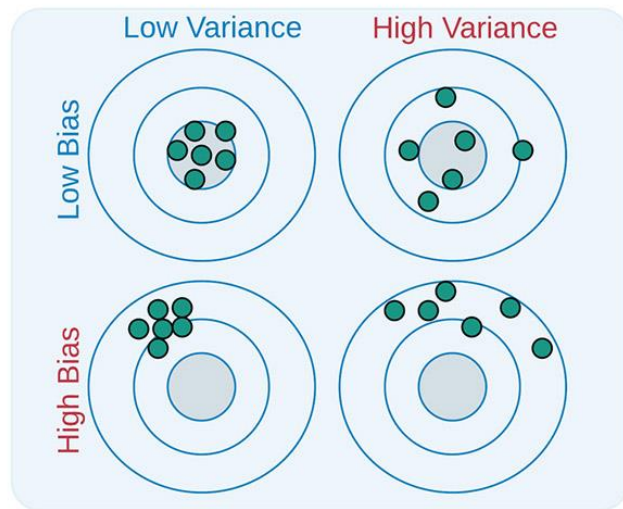
EL **Overfitting** ocurre cuando debido a la **alta** complejidad de nuestro modelo se ajusta demasiado a los datos, al contrario con datos nuevos.

El **Underfitting** ocurre cuando debido a la **baja** complejidad de nuestro modelo el ajuste con los datos no refleja de manera correcta el comportamiento de nuestros datos, ya sean de entrenamiento o de prueba.



ANÁLISIS DE LOS RESULTADOS DEL MODELO: SESGO POR LA SELECCIÓN DEL MODELO Y MUESTRA DE ENTRENAMIENTO.

Para prevenir errores por el sesgo en lo modelos, se recomienda utilizar un método de selección de la muestra de entrenamiento, probar diferentes modelos y usar un proceso de optimización de hiperparámetros.

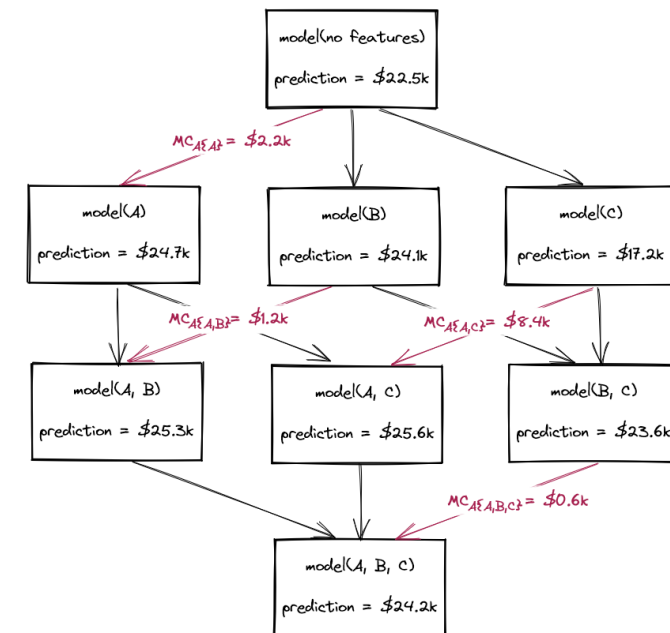
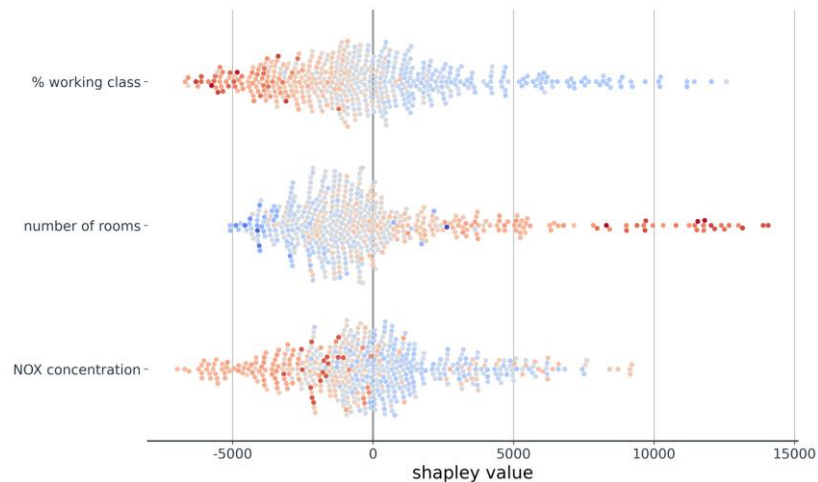


INTERPRETACIÓN DE RESULTADOS.

Uno de los pasos más importantes es la interpretación del modelo, para comprender mejor cómo las variables influyen en las predicciones.

Métodos para interpretar el modelo:

SHAP values, Feature Importance en Random Forest, visualización de pesos en redes neuronales.



EJEMPLOS DE APLICACIONES

Mantenimiento Predictivo.

Identificación de defectos en líneas de producción mediante visión artificial.

Predicción de demanda y optimización de logística.

Personalización de contenido en plataformas digitales.

Detección de actividades de riesgo y fraudes en tiempo real

Análisis de imágenes médicas y datos clínicos para detección de enfermedades.

Análisis de datos meteorológicos, humedad del suelo y condiciones ambientales para predecir el rendimiento de los cultivos

Predicción de contaminantes.

OTRAS LÍNEAS DE INVESTIGACIÓN DE INTERÉS.

Computo distribuido y procesamiento en paralelo.

Se centra en el diseño y optimización de sistemas que dividen tareas computacionales en múltiples nodos o procesadores para mejorar el rendimiento, Áreas de Aplicación: Supercomputación en ciencia e ingeniería, Big Data y procesamiento masivo de datos.

Agentes inteligentes.

Sistemas autónomos capaces de percibir su entorno, tomar decisiones y aprender para alcanzar objetivos específicos, Áreas de Aplicación: automatización de procesos, Sistemas de asistencia inteligente y toma de decisiones.

OTRAS LÍNEAS DE INVESTIGACIÓN DE INTERÉS.

Minería de procesos.

Se enfoca en el análisis de registros de eventos para descubrir, monitorear y mejorar procesos empresariales mediante técnicas de Machine Learning, Áreas de Aplicación: Optimización de flujos de trabajo en empresas.

Representación de conocimiento.

Se centra en la manera en que la información y el conocimiento pueden ser estructurados, almacenados y utilizados por sistemas inteligentes para la toma de decisiones. Áreas de Aplicación: Ontologías y bases de conocimiento en inteligencia artificial.

GRACIAS...

Datos de Contacto



Rodrigo Domínguez García
Tecnologías de la Información
Av. Miguel de Cervantes #120
Complejo Industrial Chihuahua
Chihuahua, Chih. México. C.P. 31136
Teléfono: (614) 4391154
rodrigo.dominguez@cimav.edu.mx