INTRODUCTION TO MACHINE LEARNING

Supervised Learning Algorithms

WORKSHOP DETAILS

- Instructor: Tanya Khanna
- Email: tk759@scarletmail.rutgers.edu
- Course Materials: Github Link - https://github.com/Tanya-Khanna/Data-Science-Workshop---Spring-2025---NBL-
- Workshop Recordings: https://libguides.rutgers.edu/datascience/python
- Workshop Feedback Form: https://rutgers.libwizard.com/f/graduate_specialist_feedback

WORKSHOPS SCHEDULE

Introduction to Python Programming	February 3, 2025; 2 - 3:30 PM
Mastering Data Analysis: Pandas and Numpy	February 10, 2025; 2 - 3:30 PM
Introduction to Tableau: Visualizing Data Made Easy	February 17, 2025; 2 - 3:30 PM
Introduction to Machine Learning: Supervised Learning	February 24, 2025; 2 - 3:30 PM
Introduction to Machine Learning: Unsupervised Learning	March 3, 2025; 2 - 3:30 PM
Data-Driven Decision Making: A/B Testing and Statistical Hypothesis Testing •••	March 10, 2025; 2 - 3:30 PM
Demystifying Generative AI	March 24, 2025; 2 - 3:30 PM
Large Language Models: From Theory to Implementation	March 31, 2025; 2 - 3:30 PM
Generative AI Applications with AI Agents	April 7, 2025; 2 - 3:30 PM
Building Intelligent Recommendation Systems	April 14, 2025; 2 - 3:30 PM

https://libcal.rutgers.edu/calendar/nblworkshops?cid=4537&t=d&d=0000-00-00&cal=4537&inc=0

TABLE OF CONTENTS

DECODING DATA SCIENCE, AI, ML, AND STATISTICS

Imagine you're at a large family gathering, with Data Science, Artificial Intelligence (AI), and Machine Learning (ML), Statistics as four relatives who are closely related but each with their own distinct personalities and interests.

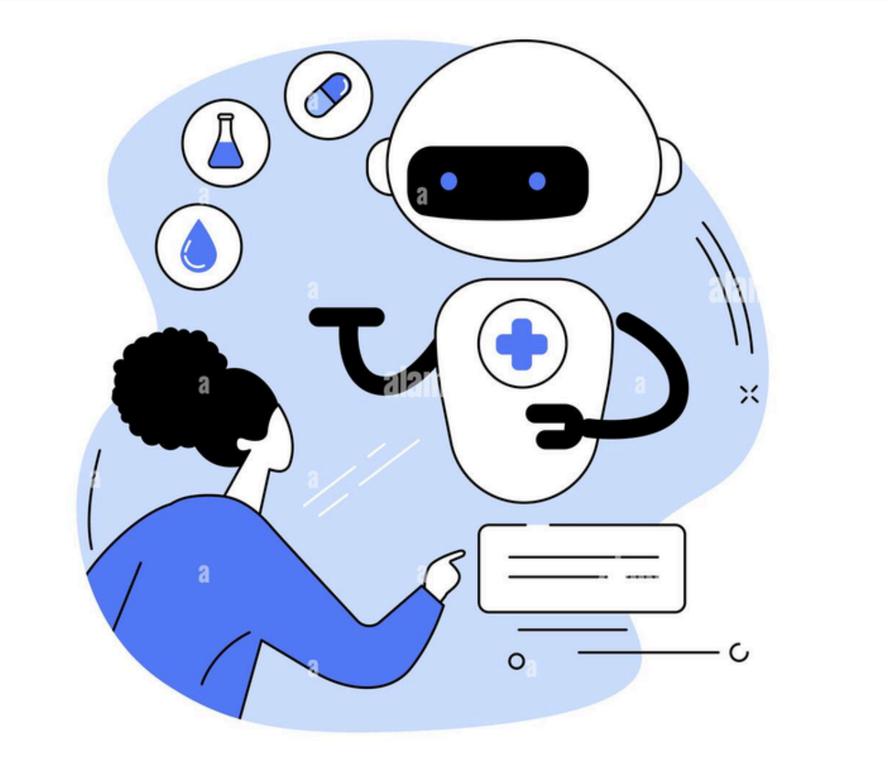
Artificial Intelligence (AI)	Machine Learning (ML)	Data Science (DS)	Statistics
AI is like the imaginative inventor in the family, always dreaming of creating machines that can think, learn, and make decisions like humans.	ML is the family's quick learner, capable of picking up patterns from data and improving over time, much like a child learning from experience.	DS is the detective with a keen eye for detail, sifting through vast amounts of data to uncover hidden insights and tell compelling stories.	Statistics is the wise elder, deeply knowledgeable about the right ways to gather, analyse, and interpret data to make informed decisions.
AI encompasses the development of computer systems that can perform tasks requiring human intelligence, including speech recognition, decision-making, and language translation.	ML is a subset of AI focused on algorithms that enable computers to learn from and make predictions or decisions based on data, without being explicitly programmed for each task.	DS combines domain expertise, programming skills, and knowledge of mathematics and statistics to extract meaningful insights from data, which can then inform decision-making and strategic planning.	Statistics is the branch of mathematics dealing with data collection, analysis, interpretation, and presentation, providing methods for making sense of numerical data and making decisions under uncertainty.
An AI system controls a self-driving car, navigating through traffic and making real-time decisions based on sensor data.	A streaming service uses ML to analyze your watching habits and recommends movies and TV shows you might like.	A retailer uses data science to analyze customer purchase histories and demographic data to optimize inventory and tailor marketing campaigns.	A researcher uses statistical methods to test the effectiveness of a new drug compared to an existing one, based on data from clinical trials.

BRIEF HISTORY OF MACHINE LEARNING

The journey of machine learning (ML) is a tale of innovation and discovery, marked by key milestones:

- ***1950s: Beginnings*** *- Alan Turing questions if machines can think, leading to early ML programs like Arthur Samuel's checkers-playing algorithm, showcasing machines learning from experience.*
- ***1960s: Early Models*** *- The perceptron, an early neural network model, is introduced by Frank Rosenblatt, laying foundational concepts despite its limitations highlighted by Minsky and Papert.*
- ***1980s: Revival*** *- Interest in ML rekindles with advances in algorithms and computational power, enabling more complex models like multi-layer neural networks.*
- ***1990s: Expansion*** *- ML applications grow, powered by new algorithms like Support Vector Machines (SVMs), improving performance in tasks like handwriting recognition.*
- ***2000s-Present: Deep Learning Era*** *- The advent of big data and improved computing capabilities leads to the rise of deep learning, transforming fields like natural language processing and computer vision, and firmly establishing ML's role in modern technology.*

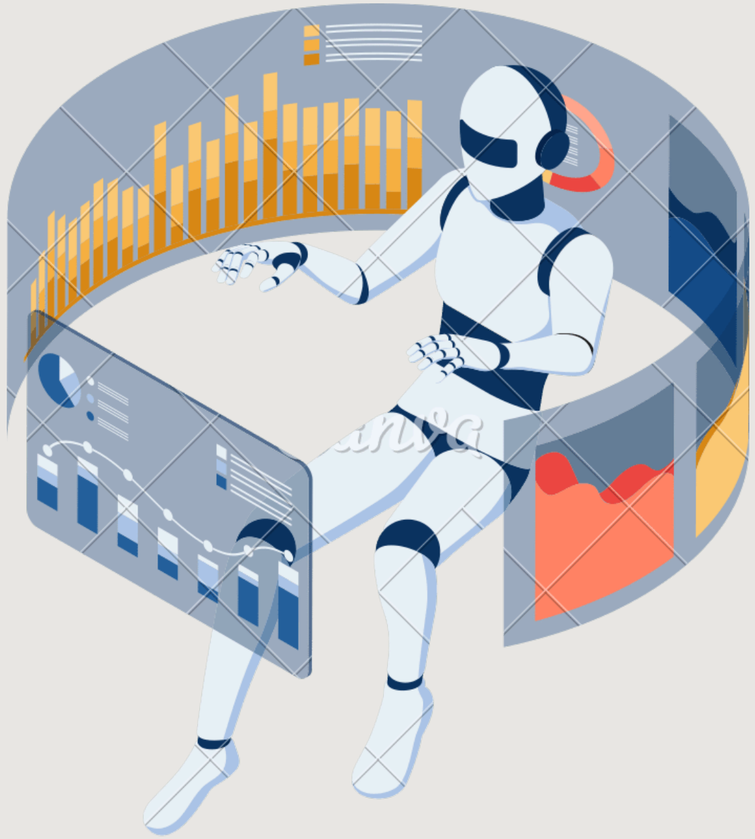
APPLICATIONS OF ML IN VARIOUS INDUSTRIES



Healthcare

- ***Disease Diagnosis and Prediction:*** *ML models can analyze medical images, genetic data, and patient history to diagnose diseases early and accurately, and predict health risks.*
- ***Personalized Medicine:*** *Tailoring treatment plans to individual genetic profiles, improving the effectiveness of treatments and reducing side effects.*
- ***Drug Discovery and Development:*** *Accelerating the search for new drugs by predicting molecule behavior and drug-target interactions.*

APPLICATIONS OF ML IN VARIOUS INDUSTRIES



Finance

- ***Fraud Detection:*** *Analyzing transaction patterns to identify and prevent fraudulent activities in real-time.*
- ***Credit Scoring:*** *Improving the accuracy of creditworthiness assessments, leading to more informed lending decisions.*
- ***Algorithmic Trading:*** *Using ML to analyze market data and execute trades at optimal times, maximizing returns.*

APPLICATIONS OF ML IN VARIOUS INDUSTRIES

Retail

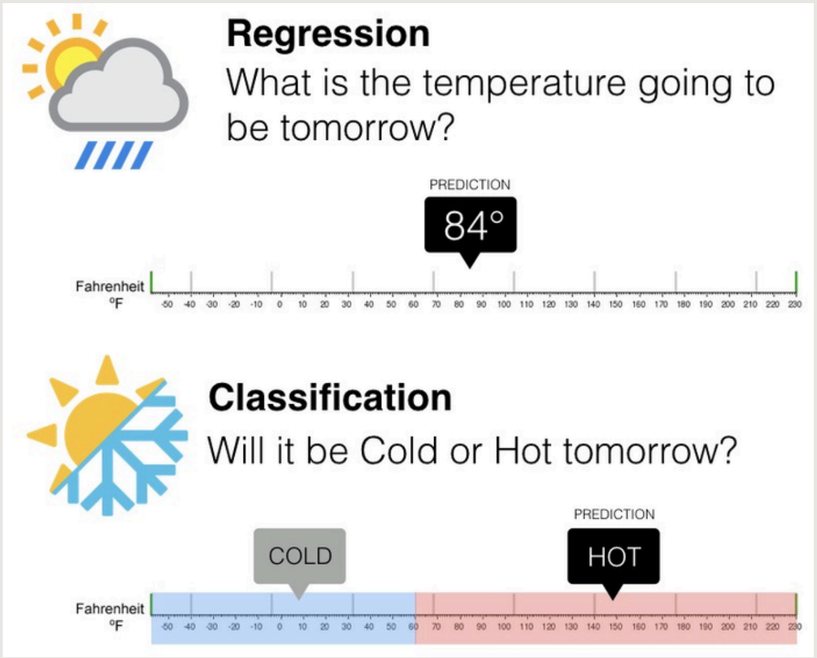
- ***Customer Segmentation and Personalization:*** *Analyzing customer behavior to provide personalized shopping experiences and recommendations.*
- ***Inventory Management:*** *Optimizing stock levels based on predictive analytics, reducing waste and ensuring product availability.*
- ***Demand Forecasting:*** *Predicting future product demand to inform production and distribution strategies.*

OVERVIEW OF TYPES OF ML

Supervised Learning	Unsupervised Learning	Reinforcement Learning
Like a student learning under the guidance of a teacher who corrects their homework. The student learns to make fewer mistakes over time.	Like exploring a treasure map without clear instructions, finding patterns and clustering similar items together on your own.	Like training a pet with rewards and penalties, where the pet learns to perform tricks to receive treats.
Involves learning a function that maps an input to an output based on example input-output pairs. It uses labeled data to predict the label for new, unseen data.	Involves learning the structure of data without labels, finding patterns or inherent structures in the input data through methods like clustering or dimensionality reduction.	An algorithm learns to perform an action within an environment to maximize some notion of cumulative reward. It's based on trial and error, using feedback from its own actions and experiences.
Predicting house prices based on features like location and size of the house. The algorithm is trained with data of houses whose prices are known.	Grouping customers based on their purchasing history without prior knowledge of the groups. The algorithm identifies patterns and clusters customers with similar behaviors.	A video game AI that learns to navigate and complete levels by trying different paths and learning from successes and failures. The AI gets "rewarded" for actions that lead to success.

TYPES OF SUPERVISED LEARNING

- ***Regression:*** *A supervised machine learning task that involves predicting continuous outcomes.*
- ***Classification:*** *A supervised machine learning task focused on predicting discrete labels or categories.*



CHARACTERISTICS OF SUPERVISED LEARNING

- ***Labeled Data:*** *The training data consists of input-output pairs, where each input is associated with a correct output (label), serving as the "answer key" for learning.*
- ***Model Training:*** *The process involves adjusting the parameters of a model to minimize the difference between the predicted outputs and the actual labels in the training data.*
- ***Generalization:*** *The ultimate goal is for the model to not only perform well on the training data but to generalize well to new, unseen data.*

KEY TERMINOLOGY IN SUPERVISED LEARNING

- ***Model:***
  - *In supervised learning, a model is essentially a mathematical formula that makes predictions. Think of it as a prediction machine that's been tuned to expect certain kinds of data (features) and use those to make guesses about something (the target). For instance, a model might predict the price of a house based on its size and location.*

- ***Algorithm:***
  - *This is the method or process used to create and train the model. It's like a recipe that tells you how to adjust your prediction machine so it gets better at making guesses. An example is the gradient descent algorithm, which fine-tunes the model by gradually reducing errors in predictions.*

- ***Features:***
  - *These are the input variables the model uses to make predictions. If our task is to predict house prices, features might include things like the number of bedrooms, square footage, or age of the house.*

KEY TERMINOLOGY IN SUPERVISED LEARNING

- ***Target:***
  - *The target is what you're trying to predict. In our house price example, the target is the actual price of the house.*

- ***Training Data:***
  - *This is the dataset used to train your model. It includes both the features (input data) and the target (what you want to predict). You can think of it as the textbook from which your model learns.*

- ***Test Data:***
  - *After training, you need to check how well your model has learned. Test data is like the final exam for your model, consisting of new examples it hasn't seen before to predict. The model's performance here gives you an idea of how well it can apply what it learned to real-world data.*

KEY TERMINOLOGY IN SUPERVISED LEARNING

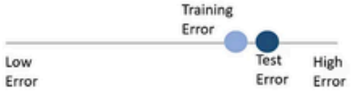
- ***Fit:***
  - *Imagine trying on clothes to find the one that suits you best; it's neither too tight nor too loose. In the world of data, "fit" refers to how well a model's predictions match up against the actual data. It's about finding the model that "fits" the data just right. The goal is to achieve a model that accurately captures the underlying pattern of the data without capturing the random noise.*
- ***Training Error:***
  - *Training error is the measure of the model's prediction error over the training dataset. It represents how well the model fits the data it was trained on, typically calculated as the average difference between the predicted values and the actual values in the training set.*
- ***Test Error:***
  - *Test error, also known as generalization error, measures the model's prediction error over a new, unseen dataset. It's indicative of how well the model can generalize from the training data to predict outcomes in unseen data, emphasizing the model's effectiveness in practical applications.*

KEY TERMINOLOGY IN SUPERVISED LEARNING

- ***Bias:***
  - *If you consistently miss basketball shots to the left, you're biased towards the left. In modeling, bias is about consistently missing the mark in a certain direction because the model makes assumptions that don't quite match up with the real world. Bias refers to the error introduced by approximating a real-world problem, which may be complex, with a simpler model. In the context of machine learning, it's the difference between the average prediction of the model and the true value that we're trying to predict. High bias can cause the model to miss relevant relations between features and target outputs (underfitting).*
- ***Variance:***
  - *If your basketball shots are all over the place, you have high variance. In modeling, variance is about how much your model's predictions change if you use a different subset of the data. It's like being inconsistent in your game. Variance measures how much the predictions of a model change if it were trained on a different dataset. It reflects the sensitivity of the model to small fluctuations in the training set. High variance can lead to models that fit the training data very well but fail to generalize to new data (overfitting).*

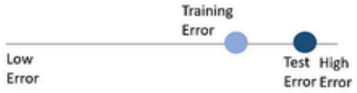
High Bias - Low Variance

Symptom
Training error is very high and Test error is almost same as Training error.



High Bias - High Variance

Symptom
Training error is very high and Test error is even higher than the Training error.



Low Bias - Low Variance

Symptom
Training error is low and so is the Test error. This is the best case scenario !

Low Bias - High Variance

Symptom
Training error is low and Test error is high. Training error is usually much lower than the Test error.

High Bias - Low Variance

Cause
- Model is underfitting and is too simple to capture the true relationship between target and predictor variables. This becomes a source of high bias.
- Simple models tend to be more stable model with low variance to change in training data.

Remedy
- Build a more complex model – Add more features, build bigger networks with more hidden nodes and layers (deep leaning), deeper trees (random forest), more trees (GBM)

High Bias - High Variance

Cause
- Model is underfitting and is too simple to capture the true relationship between target and predictor variables. This becomes a source of high bias.
- If model has limited samples to learn from this will contribute to high variance and leads to unstable model that can change with slightest change in the training data.

Remedy
- Build a more complex model – Add more features, Build bigger networks with more hidden nodes and layers (deep leaning), deeper trees (random forest), more trees (GBM)
- Get more samples in training data.

Low Bias - Low Variance

Cause
Model has the right balance between bias and variance. It is able to capture the true relationship between target and predictor variables and is stable to changes in training data.

Remedy
Model is good to go !

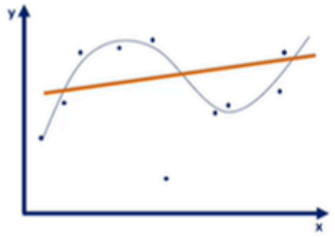
Low Bias - High Variance

Cause
- Model is overfitting to training data, it is learning both the signal and the noise in training data and does not generalize well to unknown data.
- Complex models are usually unstable and can change a lot with any change in the data.

Remedy
Build a simpler model - Hyperparameter tuning, Regularization, Feature extraction (PCA), Bagging. Get more samples in training data.

Underfitting and overfitting

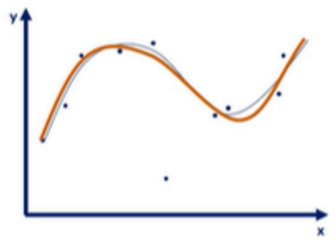
An **underfitted** model



Doesn't capture any logic

- High loss
- Low accuracy

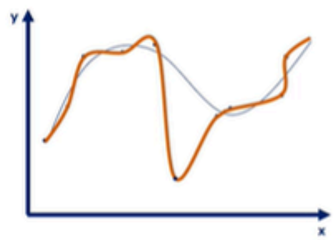
A **good** model



Captures the underlying logic of the dataset

- Low loss
- High accuracy

An **overfitted** model



Captures all the noise, thus "missed the point"

- Low loss
- Low accuracy

BIAS-VARIANCE TRADEOFF

The bias-variance tradeoff represents the balance between a model's simplicity (bias) and its sensitivity to fluctuations in the training data (variance). Striking the right balance minimizes the total error and achieves the best performance.

Bias² is the squared difference between the model's average prediction and the true values

Irreducible Error is the inherent noise in the data, which cannot be reduced by improving the model

$$\text{Total Error} = \text{Bias}^2 + \text{Variance} + \text{Irreducible Error}$$

Variance measures the model's inconsistency across different training sets

KEY TERMINOLOGY IN SUPERVISED LEARNING

- ***Parameters:***
  - *Parameters in machine learning models are the variables that the model learns from the training data. For example, weights in a neural network or coefficients in a linear regression model. These are adjusted through the learning process to minimize the model's loss function. Think of parameters as the internal settings of a machine, like the gears inside a watch. They are adjusted automatically to make sure the machine works as accurately as possible.*

- ***Hyperparameters:***
  - *Hyperparameters are the configuration settings used to structure the machine learning model. These are not learned from the data but set prior to the training process and include choices like learning rate, number of layers in a neural network, or the number of trees in a random forest. The selection of hyperparameters can significantly affect the learning process and model performance. Hyperparameters are like the settings on a machine that you can manually adjust, such as the volume on a radio. You set these before the machine starts to ensure it operates in the best possible way.*

KEY TERMINOLOGY IN SUPERVISED LEARNING

- ***Loss Function:***
  - *Imagine you're throwing darts, and the loss function is the measure of how far off you are from the bullseye. The goal is to adjust your throw to minimize this distance. A loss function, or cost function, quantifies the difference between the predicted values and the actual values in the dataset. It's a mathematical function used in the training process to guide the optimization of the model's parameters. Minimizing the loss function helps in adjusting the model to make more accurate predictions. Common examples include Mean Squared Error (MSE) for regression tasks and Cross-Entropy for classification tasks.*
- ***Evaluation:***
  - *Evaluation in machine learning involves assessing the performance of a model on a separate dataset not used during training, often called the test dataset. It gives insights into how well the model generalizes to new, unseen data. This process uses specific metrics to quantify the model's performance. Evaluation is like getting a report card that tells you how well you did in your final exams. It's about assessing the performance of your model after it has been trained, to see how well it does on unseen data.*

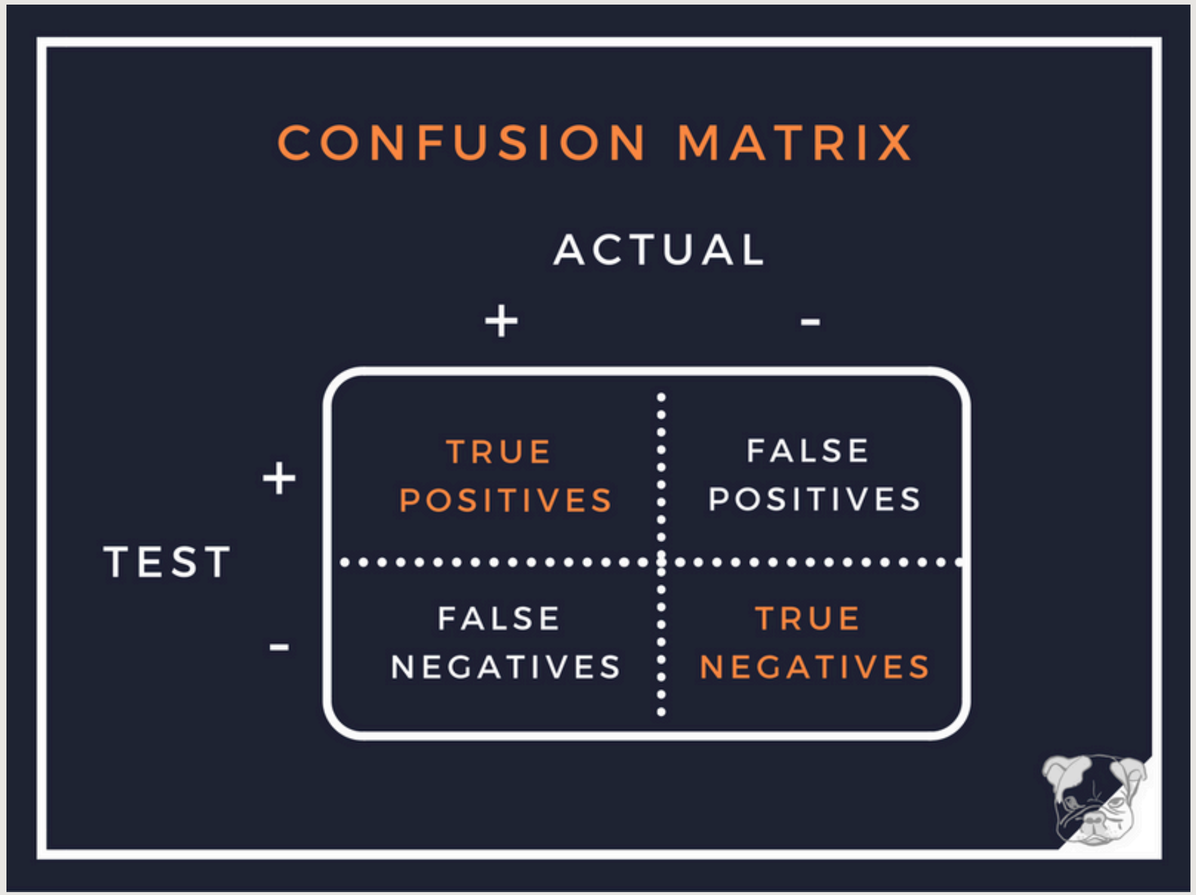
EVALUATION METRICS FOR REGRESSION

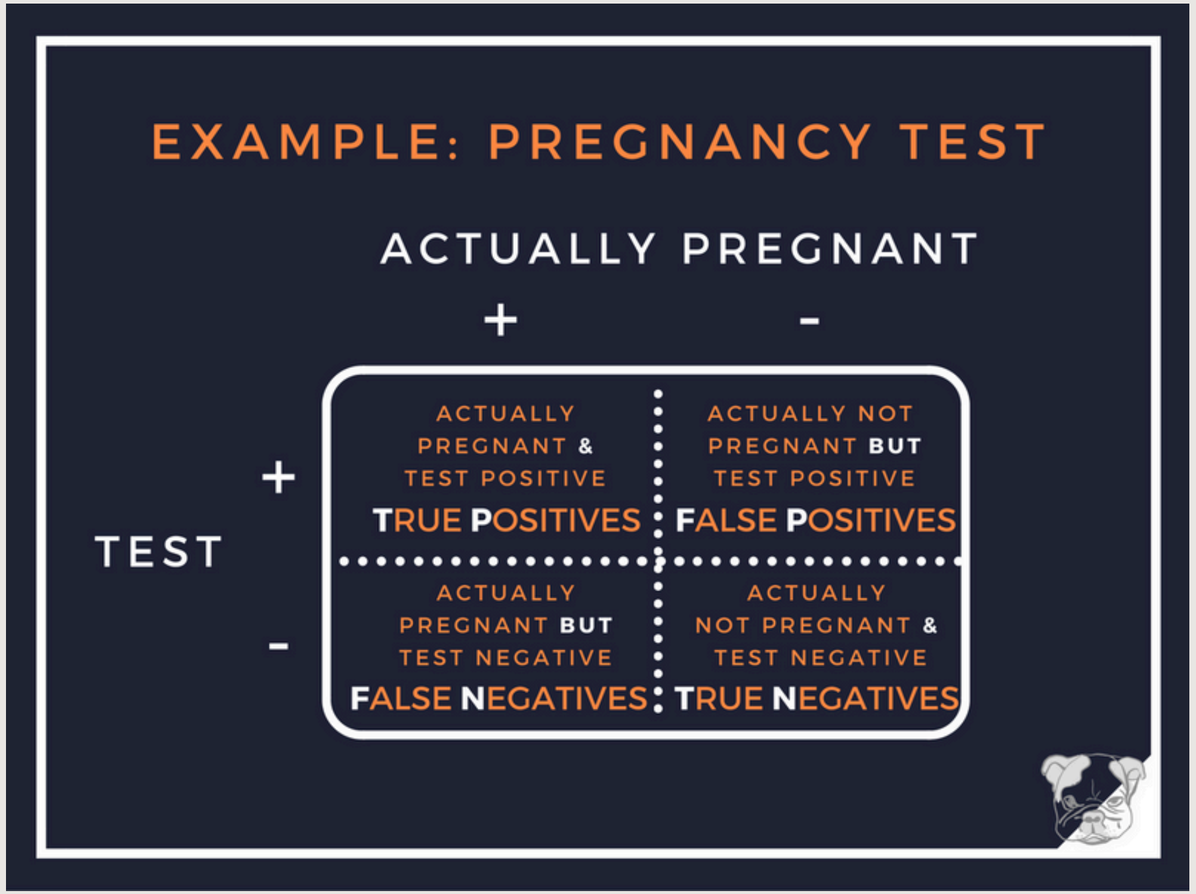
In the context of regression, where you're predicting continuous values (like house prices), metrics are like different ways of measuring your accuracy. Did you get close to the right price? Were you way off? Metrics help answer these questions. Metrics for regression tasks quantify the accuracy of the model's predictions against the actual values. Common metrics include:

- ***Mean Absolute Error (MAE):*** *The average absolute difference between predicted and actual values.*
- ***Mean Squared Error (MSE):*** *The average of the squared differences between predicted and actual values. It heavily penalizes larger errors.*
- ***Root Mean Squared Error (RMSE):*** *The square root of MSE, bringing the errors back to the original units of the output variable.*
- ***R-squared (R^2):*** *A statistical measure that represents the proportion of the variance for the dependent variable that's explained by the independent variables in the model. It ranges from 0 to 1, where 1 means the model perfectly predicts the target variable.*

Each of these metrics offers a different perspective on the model's performance, helping to identify areas for improvement or adjustment.

EVALUATION METRICS FOR CLASSIFICATION





EVALUATION METRICS FOR CLASSIFICATION

EXAMPLE: CALCULATIONS

	+	-	
+	TP = 50	FP = 10	(TP + FP) = 60
-	FN = 5	TN = 35	(FN + TN) = 40
	(TP + FN) = 55	(FP + TN) = 45	

ACCURACY | HOW ACCURATE THE CLASSIFER IS

$$\frac{TP + TN}{TP + FP + FN + TN} = \frac{50 + 35}{60 + 40} = \frac{85}{100}$$

EXAMPLE: CALCULATIONS

	+	-	
+	TP = 50	FP = 10	(TP + FP) = 60
-	FN = 5	TN = 35	(FN + TN) = 40
	(TP + FN) = 55	(FP + TN) = 45	

MISCALCULATION RATE | ERROR RATE

$$\frac{FP + FN}{TP + FP + FN + TN} = \frac{10 + 5}{60 + 40} = \frac{15}{100}$$

EVALUATION METRICS FOR CLASSIFICATION

EXAMPLE: CALCULATIONS

	+	-	
+	TP = 50	FP = 10	(TP + FP) = 60
-	FN = 5	TN = 35	(FN + TN) = 40
	(TP + FN) = 55	(FP + TN) = 45	

TRUE POSITIVE RATE | OR **SENSITIVITY/ RECALL** - HOW OFTEN A PREGNANT LADY IS TESTED PREGNANT

$$\frac{\text{TRUE POSITIVES}}{\text{ALL ACTUAL POSITIVES}} = \frac{TP}{TP + FN} = \frac{50}{55}$$

EXAMPLE: CALCULATIONS

	+	-	
+	TP = 50	FP = 10	(TP + FP) = 60
-	FN = 5	TN = 35	(FN + TN) = 40
	(TP + FN) = 55	(FP + TN) = 45	

FALSE POSITIVE RATE | HOW OFTEN A NON-PREGNANT LADY IS TESTED PREGNANT

$$\frac{\text{FALSE POSITIVE}}{\text{ALL ACTUAL NEGATIVES}} = \frac{FP}{FP + TN} = \frac{10}{45}$$

EVALUATION METRICS FOR CLASSIFICATION

EXAMPLE: CALCULATIONS

	+	-	
+	TP = 50	FP = 10	(TP + FP) = 60
-	FN = 5	TN = 35	(FN + TN) = 40
	(TP + FN) = 55	(FP + TN) = 45	

PRECISION | PROPORTION OF THE POSITIVE TESTS THAT ARE CORRECT

$$\frac{\text{TRUE POSITIVE}}{\text{ALL POSITIVE TESTS}} = \frac{TP}{TP + FP} = \frac{50}{60}$$

OTHER MEASURES

GEOMETRIC MEAN | MEASURES THE PERFORMANCE OF THE CLASSIFIER

$$\sqrt{TP \times PRECISION} \quad \text{OR} \quad \sqrt{TP \times TN}$$

F-MEASURE | 0 <= **b**, **b** REGULATES THE RELATIVE IMPORTANCE OF PRECISION WITH RESPECT TO THE TRUE POSITIVE RATE

$$\frac{(b\hat{}2 - 1)\ TP \times PRECISION}{b\hat{}2(PRECISION) + TP}$$