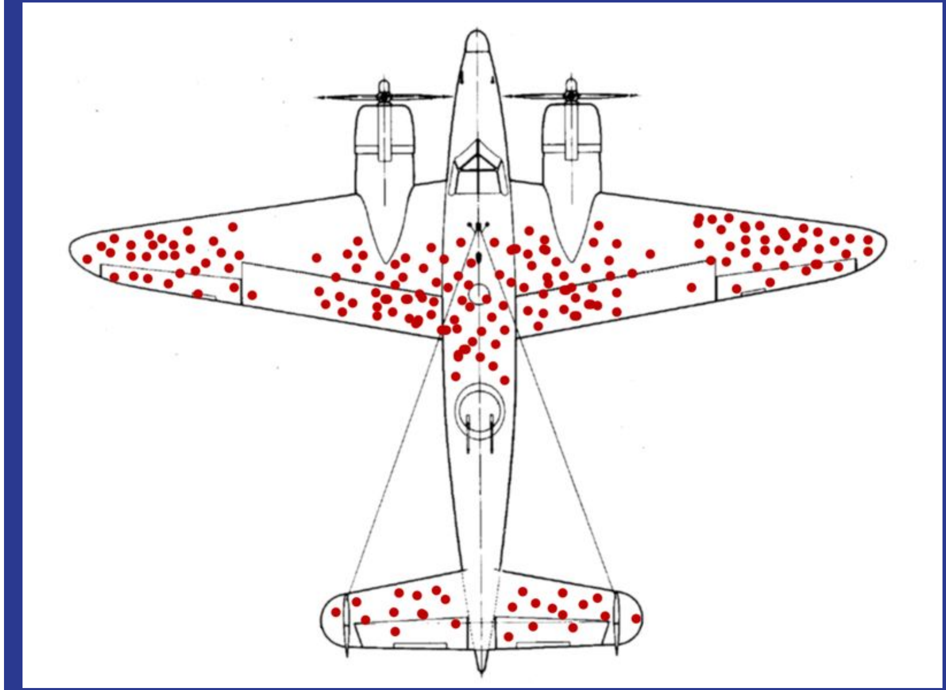
Intro to Machine Learning

Dipon Talukder
Assessment Developer (Data and AI), Workera.ai
Research Assistant, Universal Machine.io
Adjunct Faculty, East Delta University

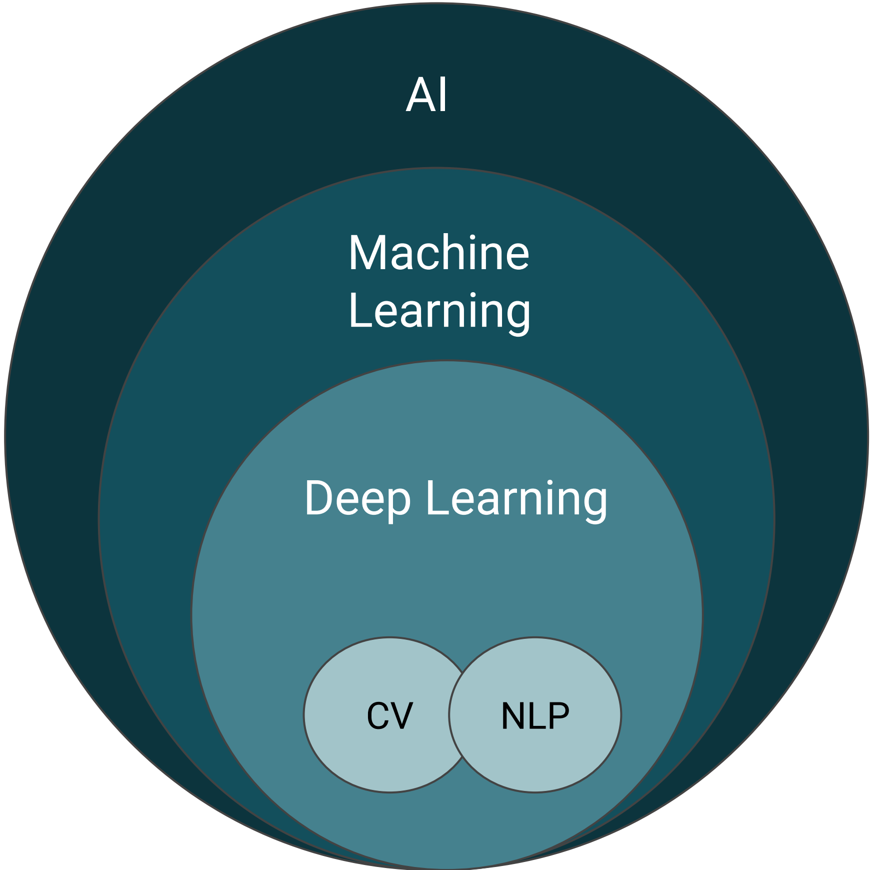
My Motivation



My Motivation

Target's Dilemma: Exposing A Teen Girl's Pregnancy

The AI Realm



AI System

- AI is computer software that mimics the ways that humans think in order to perform complex tasks, such as analyzing, reasoning, and learning.

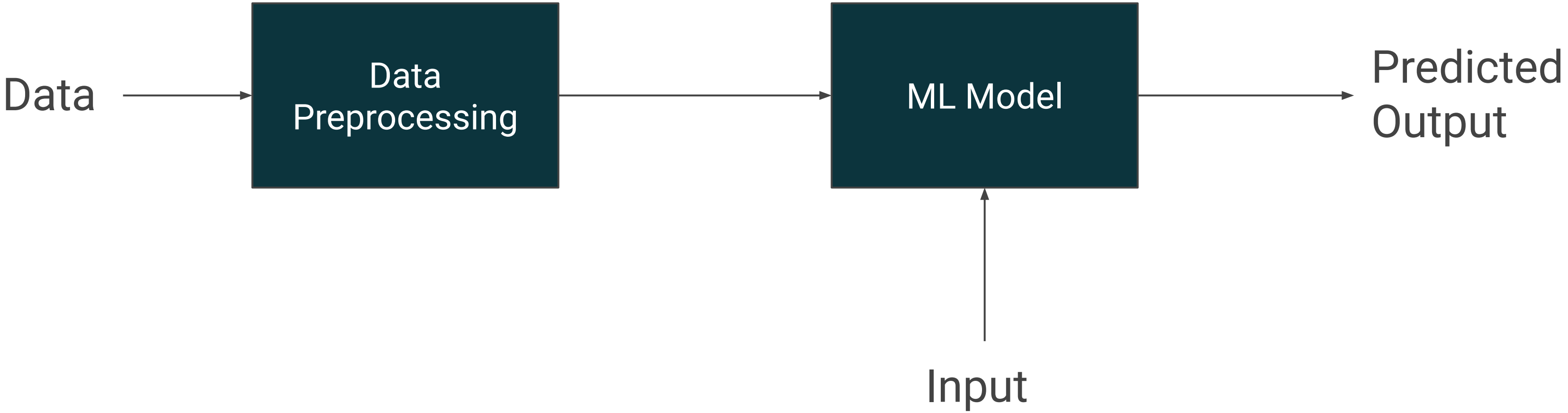
An AI System that is not a ML system):
Rule-Based Expert System: Diagnosing a Common Cold

Knowledge Base: Rules in the form of if-then statements about symptoms and diagnosis.
Inference Engine: Part of the system that applies the rules to the inputs to derive a conclusion.

- If the patient has a cough, then suspect a cold.
- If the patient has a fever, then consider flu or cold.
- If the patient has a runny nose and sneezing, then suspect a cold.
- If the patient has all three symptoms (cough, fever, and runny nose), then diagnose with a high likelihood of a cold.

Machine Learning

- ML is the science of developing algorithms and statistical models that computer systems use to perform complex tasks without explicit instructions.
- ML models are used for prediction tasks and predictions are done using the ML models.



Machine Learning

Advantages

- Simpler to Train: Requires less computational resources than deep learning for many tasks.
- Interpretability: Easier to understand and interpret the model's decisions, especially with simpler algorithms like linear regression and decision trees.
- Data Requirements: Effective with smaller datasets, unlike deep learning which typically requires large amounts of data.

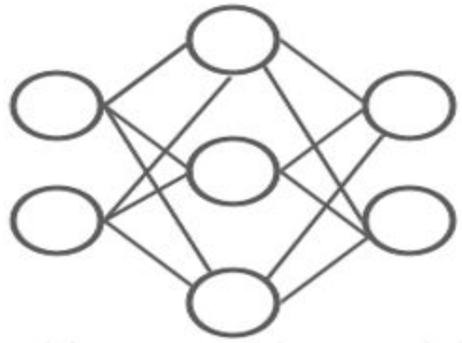
Disadvantages:

- Limited by Features: Heavily dependent on feature engineering and the quality of input data.
- Performance Plateau: Might not achieve the high accuracy that deep learning can reach on complex tasks like image and speech recognition.
- Scalability Issues: Although simpler models scale well, more complex models can become resource-intensive without significant performance gains.

Deep Learning

- A subset of machine learning where artificial neural networks, algorithms inspired by the human brain, learn from large amounts of data.
- Almost like ML but instead of traditional ML models, neural networks are used as models.
- Models are often referred as architectures.

Data/Input



Predicted Output

What are the differences?

Deep Learning

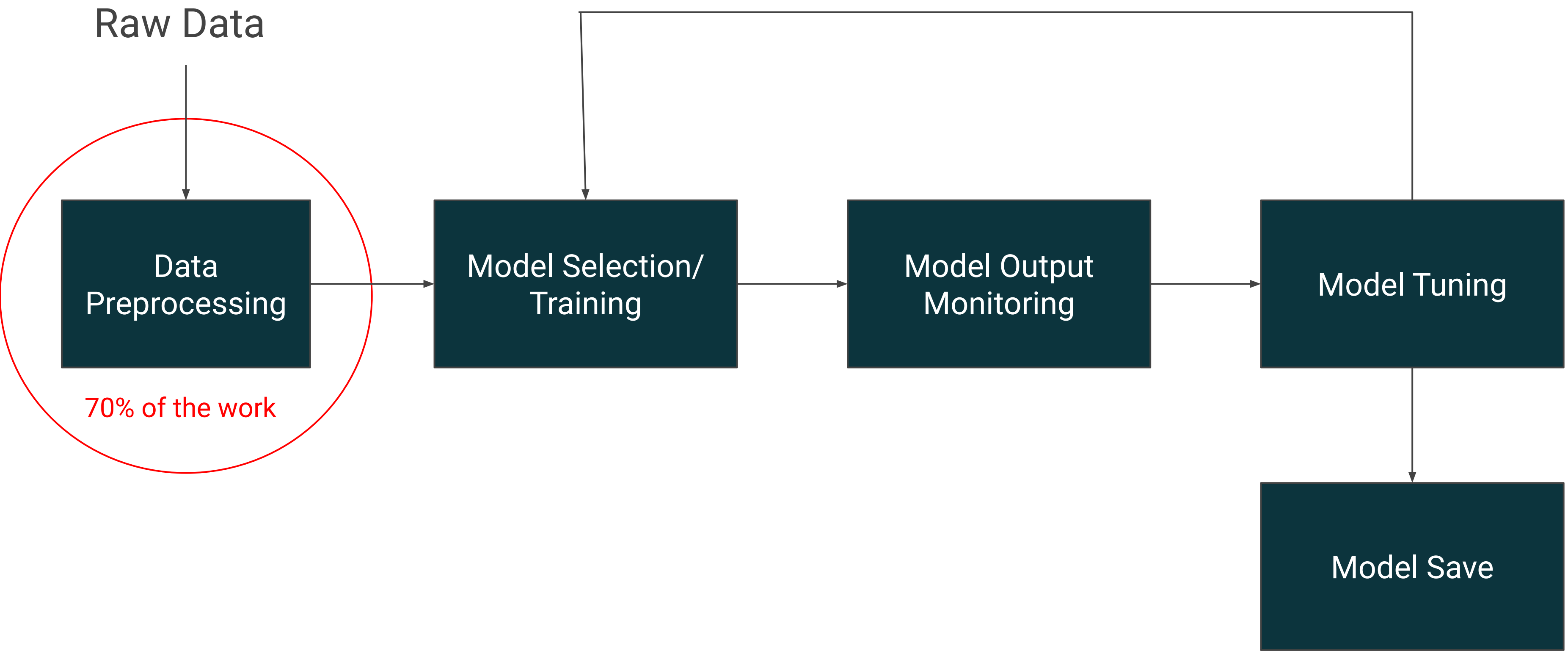
Advantages:

- Automated Feature Extraction: Does not require manual extraction of features, significantly reducing the effort and increasing the effectiveness of the algorithm.
- Scalability: Continuously improves performance with the addition of data.
- Versatility: Can be adapted to a wide range of industries including healthcare, finance, and entertainment.

Challenges with Deep Learning:

- Requires Extensive Data: Effective training requires vast amounts of labeled data.
- High Resource Consumption: Needs substantial computational power, often requiring GPUs for efficient processing.
- Lack of Interpretability: Models are often seen as black boxes with little transparency in decision-making processes. [Famous Google Issue/ Researchers stopped innovating in 2023 for six months]

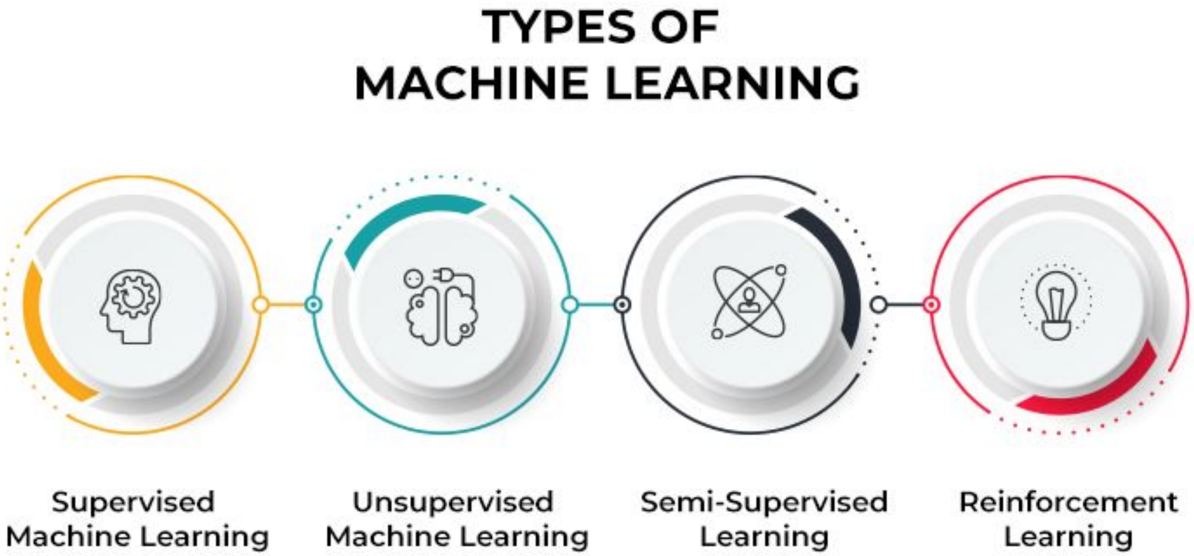
Stages of Machine Learning Model Development



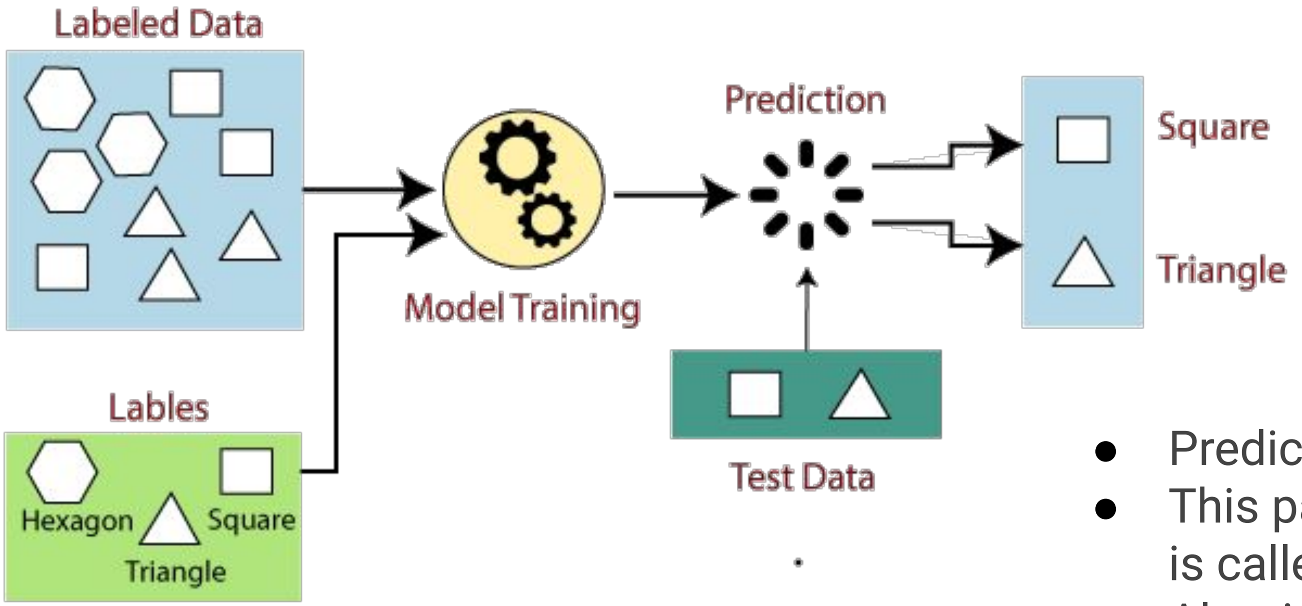
Steps Dissection

2. Model Selection

2.1 Machine Learning Models

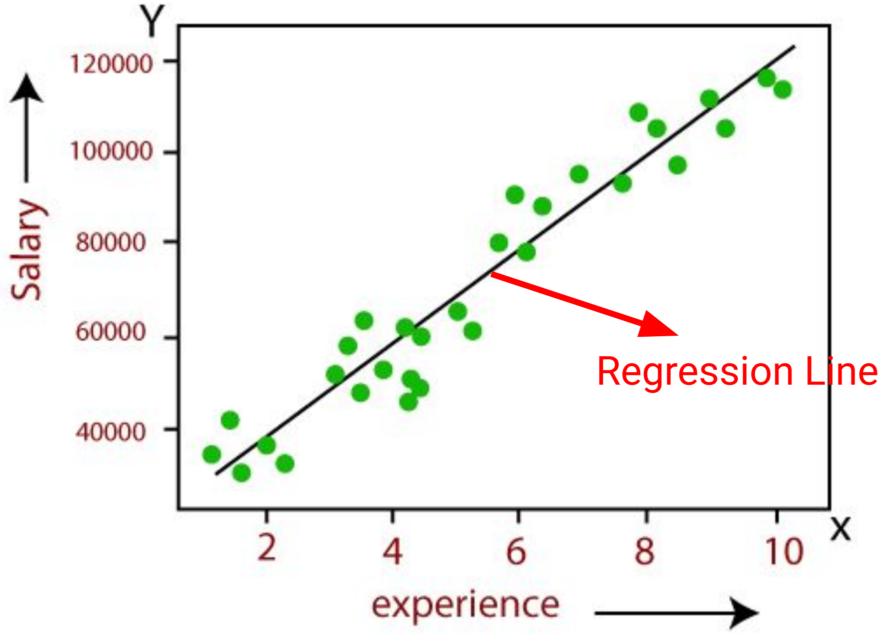


Supervised Learning



- Predicting categories
- This particular type of problem is called Classification.
- Algorithms: Random forest, XGBoost, Naive-Bayes, SVM

Supervised Learning



- Predicting numbers
- This particular type of problem is called Regression.
- Algorithms: Linear regression, Polynomial regression

Quick Quiz Time

1. A tool is being developed to identify if a patient has a specific disease, using features like symptoms, age, and previous medical conditions.

   Ans: Classification

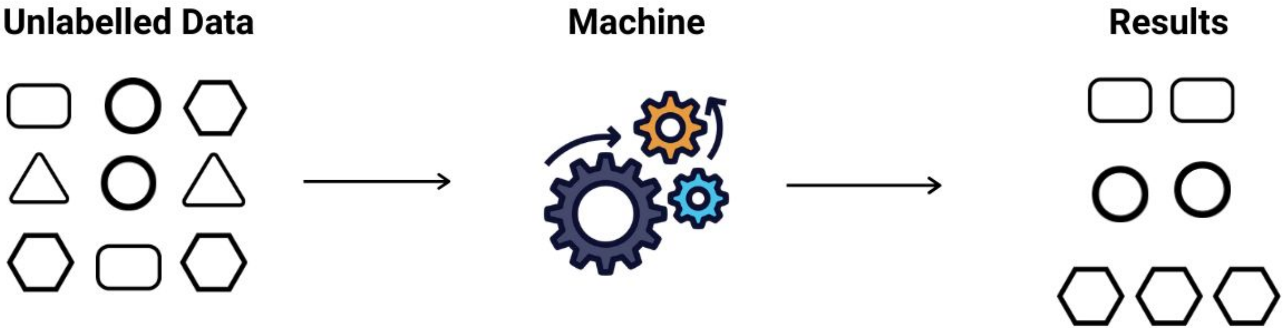
2. A company is creating a tool to estimate house prices based on features such as the house's location, square footage, number of bedrooms, and how old it is.

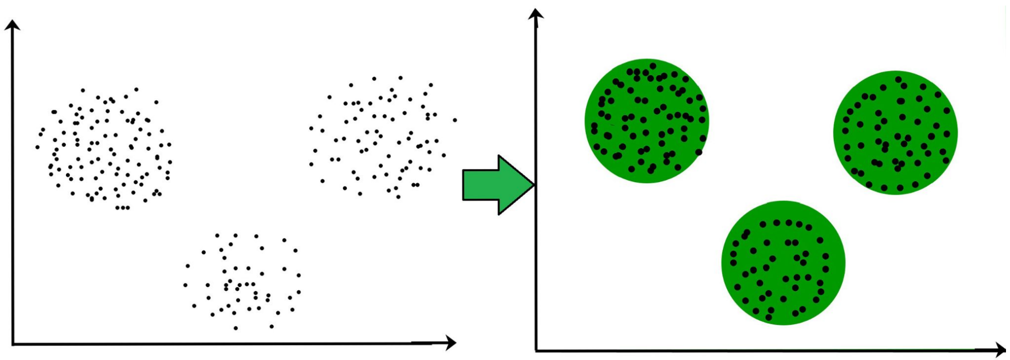
   Ans: Regression

3. A company plans to score customer happiness from 1 to 100 based on various feedback metrics, including service satisfaction, product quality, and interaction with customer support.

   Ans: Regression

Unsupervised Learning

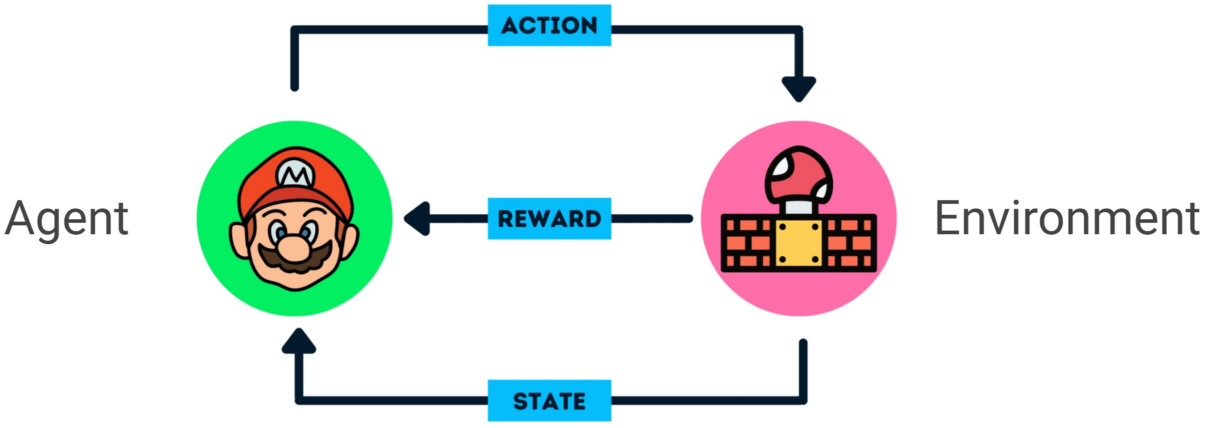




- This particular type of problem is called Clustering.

Reinforcement Learning

- A type of machine learning where an agent learns to make decisions by performing actions and receiving feedback in the form of rewards or penalties.



Quick Quiz Time

1. A marketing team wants to segment their customer base into distinct groups based on purchasing behavior to tailor marketing strategies. They have extensive data on purchase history, product preferences, and demographics but no predefined labels for the segments.

   Ans: Unsupervised

2. An email service provider aims to filter out spam emails from users' inboxes. They have a large dataset of emails labeled as "spam" or "not spam."

   Ans: Supervised

Steps Dissection

3. Model Monitoring

3.1 Performance Metrics

For classification:

- Accuracy
- Confusion Matrix
- Precision
- Recall
- F-Score

For regression:

- Mean Squared Error
- Mean Absolute Error
- R Score

Accuracy

$$Accuracy = \frac{Number\ of\ Correct\ Predictions}{Total\ number\ of\ predictions}$$

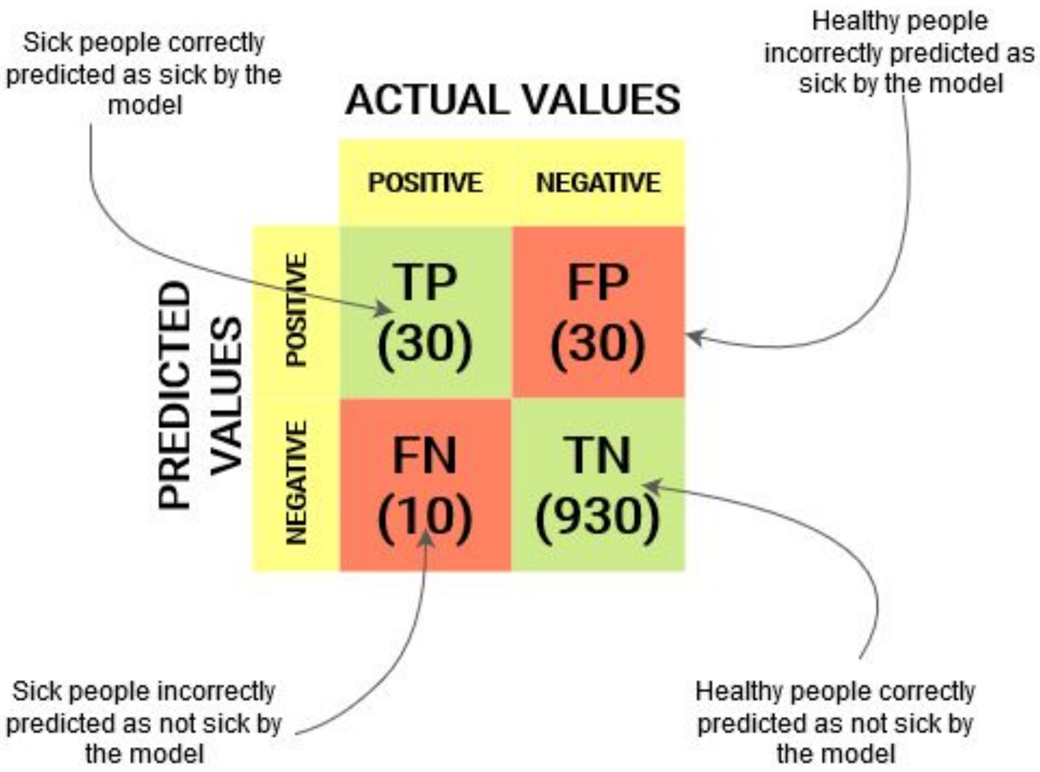
It is recommended not to use the Accuracy measure when the target variable majorly belongs to one class. For example, Suppose there is a model for a disease prediction in which, out of 100 people, only five people have a disease, and 95 people don't have one. In this case, if our model predicts every person with no disease (which means a bad prediction), the Accuracy measure will be 95%, which is not correct.

Confusion Matrix

		Predicted	
		Negative **(N)** **-**	Positive **(P)** **+**
Actual	Negative **-**	True Negative **(TN)**	**False Positive (FP)** **Type I Error**
	Positive **+**	**False Negative (FN)** **Type II Error**	True Positive **(TP)**

Confusion Matrix

Classification Problem:
Healthy or Sick?



Precision

		Predicted	
		Negative (N) -	Positive (P) +
Actual	Negative -	True Negative (TN)	**False Positive (FP)** **Type I Error**
	Positive +	**False Negative (FN)** **Type II Error**	True Positive (TP)

$$Precision = \frac{TP}{(TP + FP)}$$

- Precision is a measure of how accurate a model's positive predictions are.
- Out of all the predicted positives, how many of them are actually positive.
- Example use: Spam detection

Recall

$$\text{Recall} = \frac{TP}{TP+FN}$$

		Predicted	
		Negative (N) -	Positive (P) +
Actual	Negative -	True Negative (TN)	**False Positive (FP) Type I Error**
	Positive +	**False Negative (FN) Type II Error**	True Positive (TP)

- Recall measures the effectiveness of a classification model in identifying all relevant instances from a dataset.
- Answers the question: “Out of all the actual positives, how many of them are correctly predicted?”
- Example use: Medical sector

F1 Score

		Predicted	
		Negative (N) -	Positive (P) +
Actual	Negative -	True Negative (TN)	**False Positive (FP)** **Type I Error**
	Positive +	**False Negative (FN)** **Type II Error**	True Positive (TP)

$$F1 - score = 2 * \frac{precision * recall}{precision + recall}$$

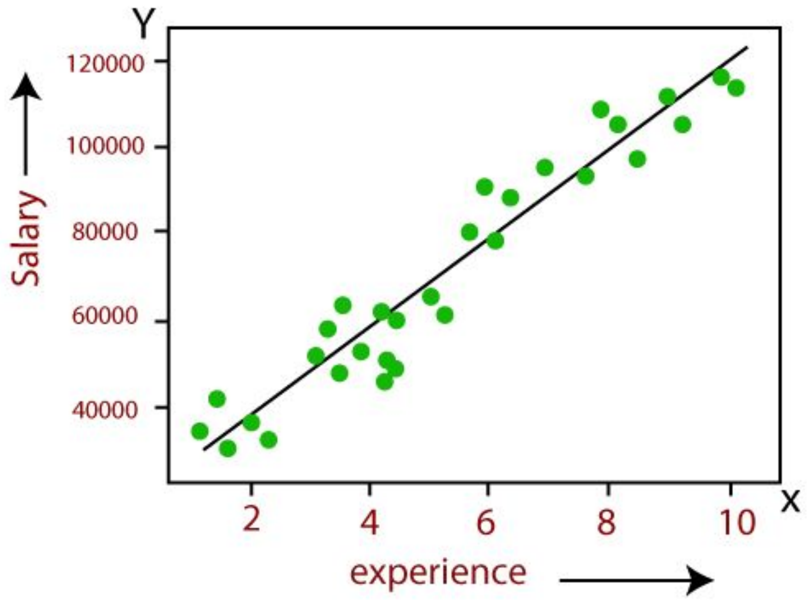
As F-score make use of both precision and recall, so it should be used if both of them are important for evaluation, but one (precision or recall) is slightly more important to consider than the other. For example, when False negatives are comparatively more important than false positives, or vice versa.

Mean Squared Error and Mean Absolute Error

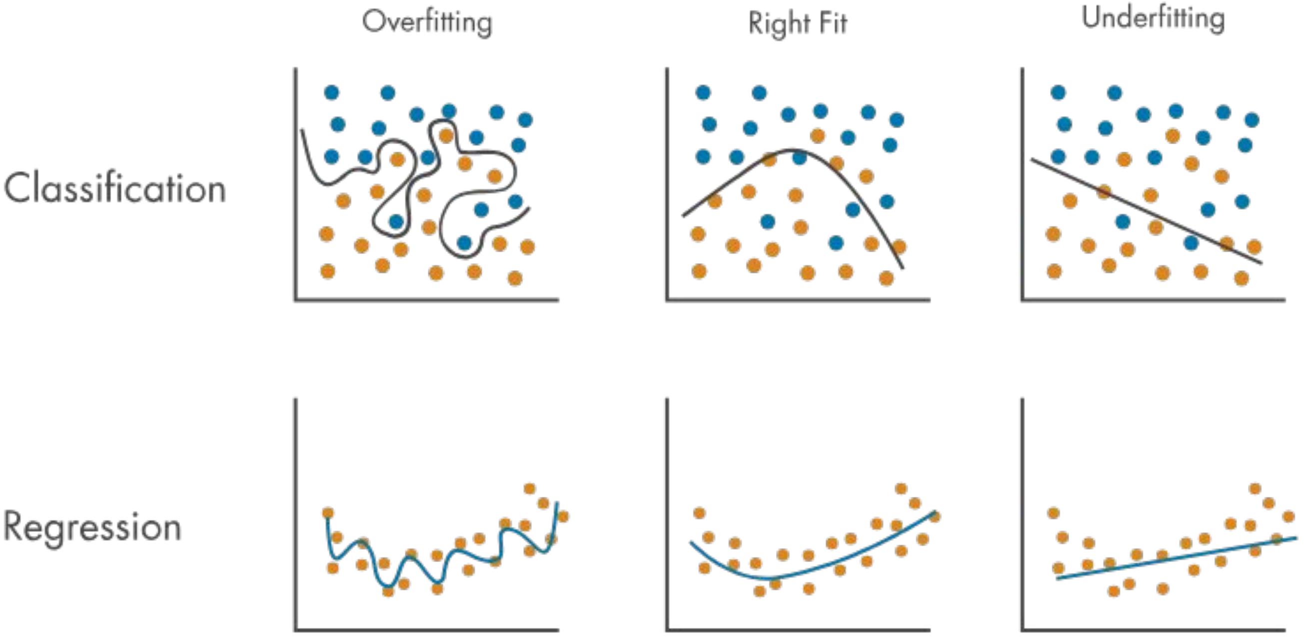
- MSE measures the average of the Squared difference between predicted values and the actual value given by the model.
- MAE measures the average of the absolute difference between predicted values and the actual value given by the model.

$$MSE = 1/N \sum (Y - Y')^2$$

$$MAE = 1/N \sum |Y - Y'|$$



3.1 Common Performance Issues(Overfit vs Underfit)

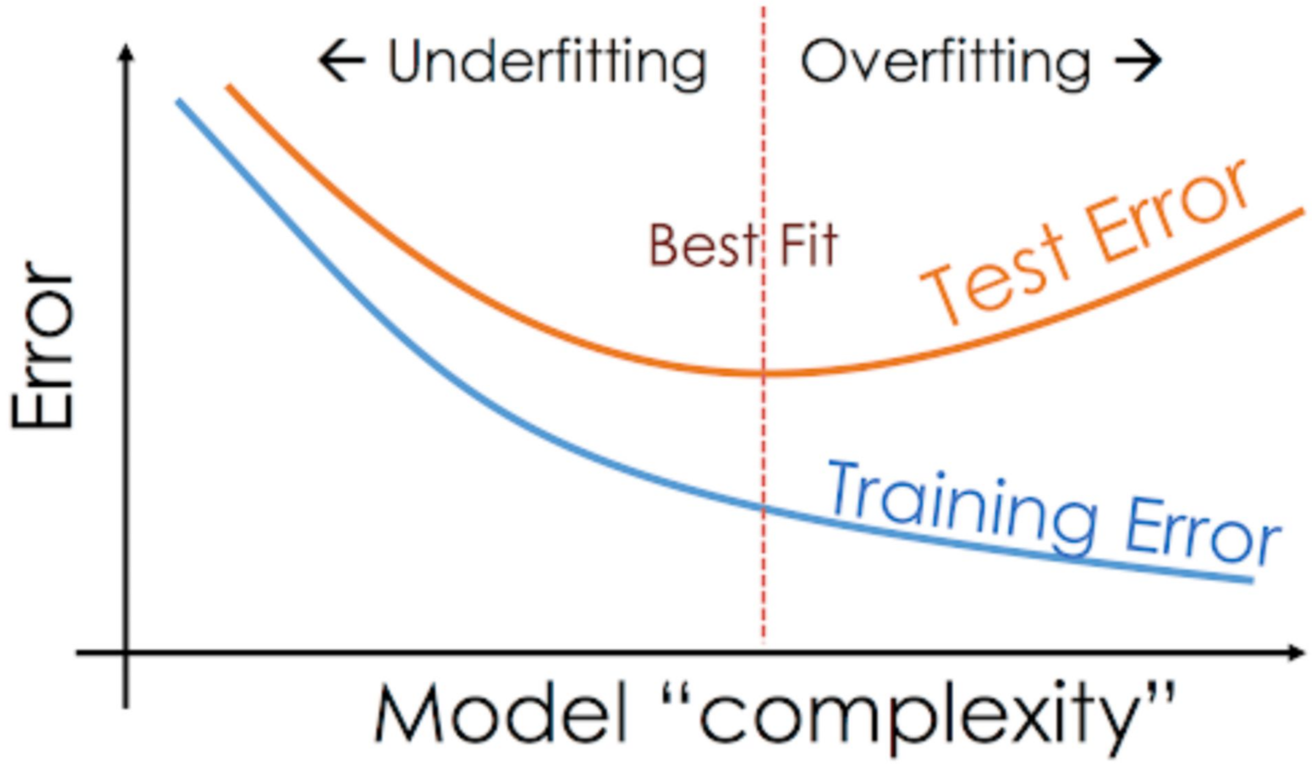


Low Bias, High Variance

Balanced

High Bias, Low Variance

3.1 Common Performance Issues(Overfit vs Underfit)



Steps Dissection

1. Data Preprocessing

1.1 Data Preprocessing

- **Data Cleaning**
  - Handling Missing Values
  - Removing Duplicates
  - Filtering Outliers
- **Data Split**
- **Data Transformation**
- **Data Encoding**
- **Advanced Feature Engineering (Optional)**

1.1.1 Handling Missing Values

- Deleting Records
  - Usually not ideal.
  - **Use When:** Data is randomly missing and comprises a small portion of the dataset.
- Interpolation
  - **Use When:** Data follows a logical sequence (e.g., time series).

Missing values

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	male	22	1	0	A/S 21171	7.25		S
2	1	1	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	female	26	0	0	STON/O2. 3101282	7.925		S
4	1	1	female	35	1	0	113803	53.1	C123	S
5	0	3	male	35	0	0	373450	8.05		S
6	0	3	male		0	0	330877	8.4583		Q

1.1.1 Handling Missing Values

- Imputation
  - Mean/Median Imputation
    - **Use When:** Data is missing at random and the missing value is a numeric variable.
  - Mode Imputation
    - **Use When:** The feature is categorical.
  - Using Models
    - **Use When:** Missing values are systematic and other variables can predict their absence.

1.2 Data Transformation

Common transformation techniques:

1. Normalization
2. Standardization

Normalization

- The main objective of normalization is to rescale the features to a standard range of values which is usually 0-1.
- Normalization is usually used when different features have different range of values and some feature might contribute more to the model learning process
- Normalization helps in equalizing the range of the features and makes sure that the features contribute equally to the learning algorithm.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Standardization

- Standardardization is also known as z-score normalization, the objective of standardization is to transform the feature such that the value of mean becomes 0 and the value of standard deviation becomes 1.
- Standardization is usually useful when features have different scales but follow normal distribution, it helps machine learning algorithms which relies on gradient based optimization to converge at a faster rate.

$$Z = \frac{\overbrace{x}^{\text{Score}} - \overbrace{\mu}^{\text{Mean}}}{\underbrace{\sigma}_{\text{SD}}}$$

1.3 Data Encoding

Three common methods:

1. Label Encoding
2. Ordinal Encoding
3. One Hot Encoding

Label encoding

- Each category is assigned a unique integer label.
- Cons: Machine learning algorithms may misinterpret the integer labels as having mathematical significance
- When to use: Categorical features with two categories

	Size	Size_encoded
0	Small	2
1	Medium	1
2	Large	0
3	Medium	1
4	Small	2

Ordinal encoding

- Similar to label encoding
- The only advantage is it allows you to explicitly define the mapping between categories and integer labels.

	Size	Size_encoded
0	Small	2
1	Medium	1
2	Large	0
3	Medium	1
4	Small	2

I want Small to be 0, Medium to be 1, and Large to be 2. It is possible in ordinal encoding but not in label encoding.

One hot encoding

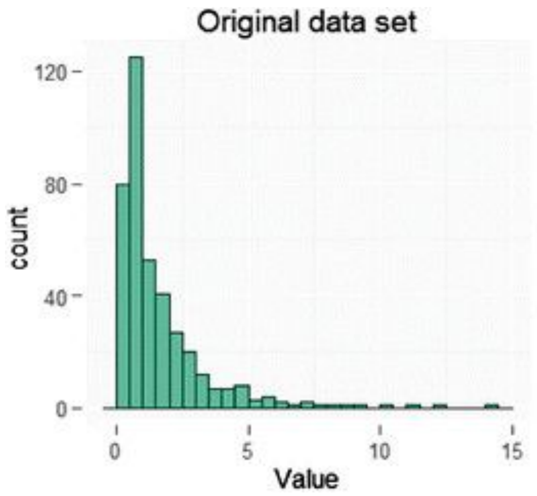
- It is suitable where the categories have no inherent order or relationship.
- For each category in a categorical column, a new binary column is created
- The binary column will have a value of 1 if the class is present, else it will be zero
- **Cons:** It can lead to a high dimensionality problem when dealing with a large number of classes

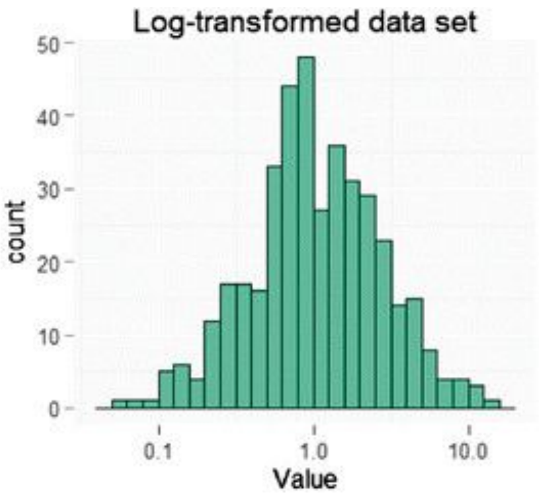
Color	Color_Yellow	Color_Blue	Color_Green
Yellow	1	0	0
Blue	0	1	0
Green	0	0	1

1.4 Advanced Feature engineering (optional)

Data distribution transformation:

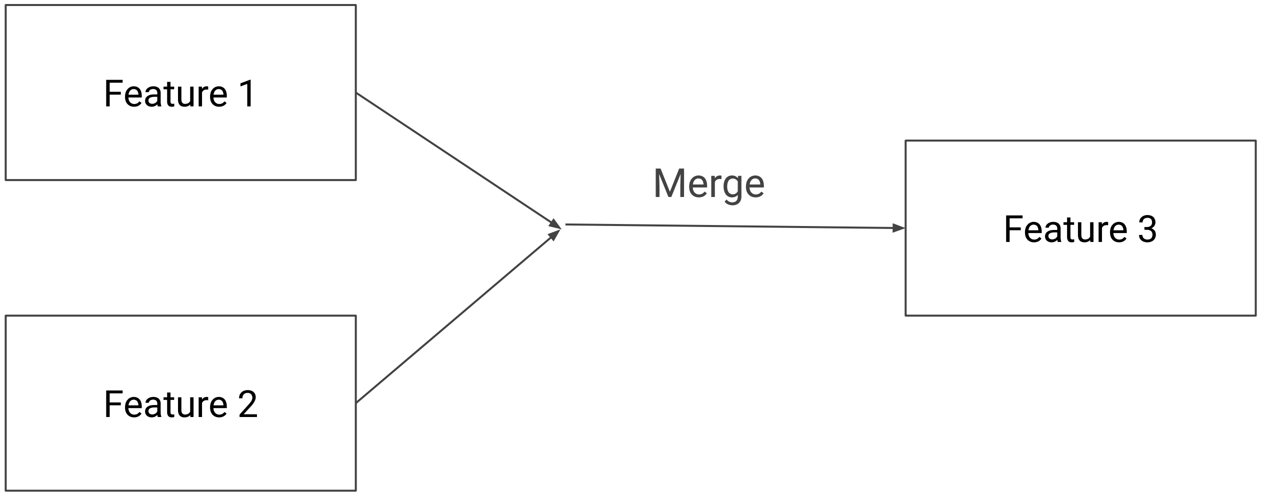
1. Log transformation
2. Box-cox transformation





1.4 Advanced Feature engineering (optional)

Feature creation



Thank You