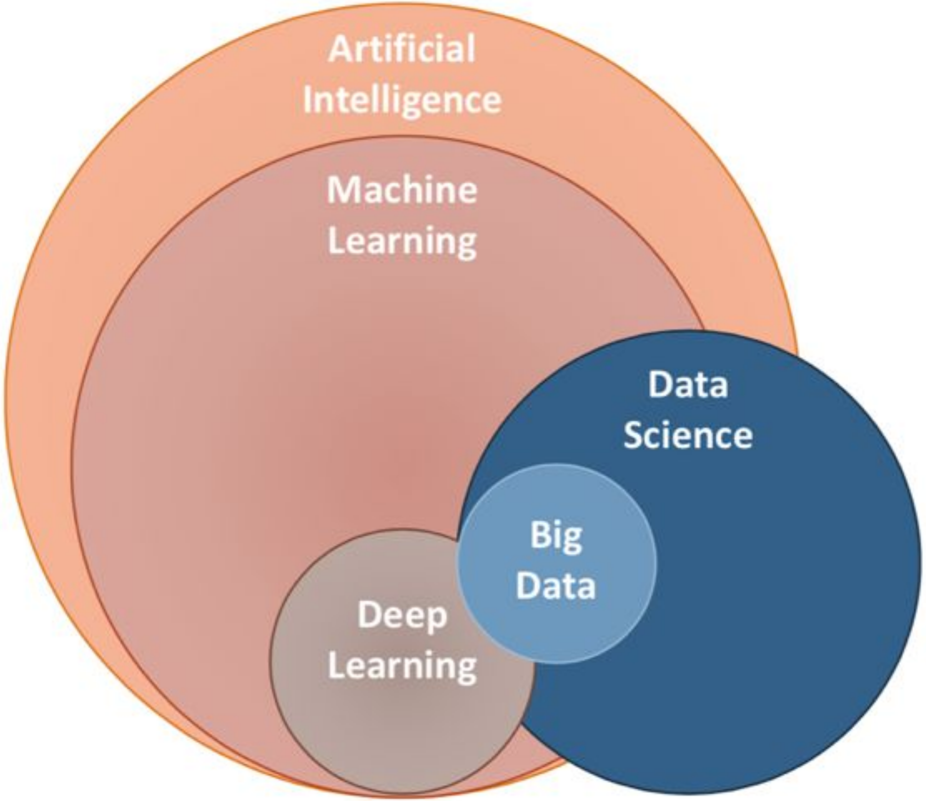
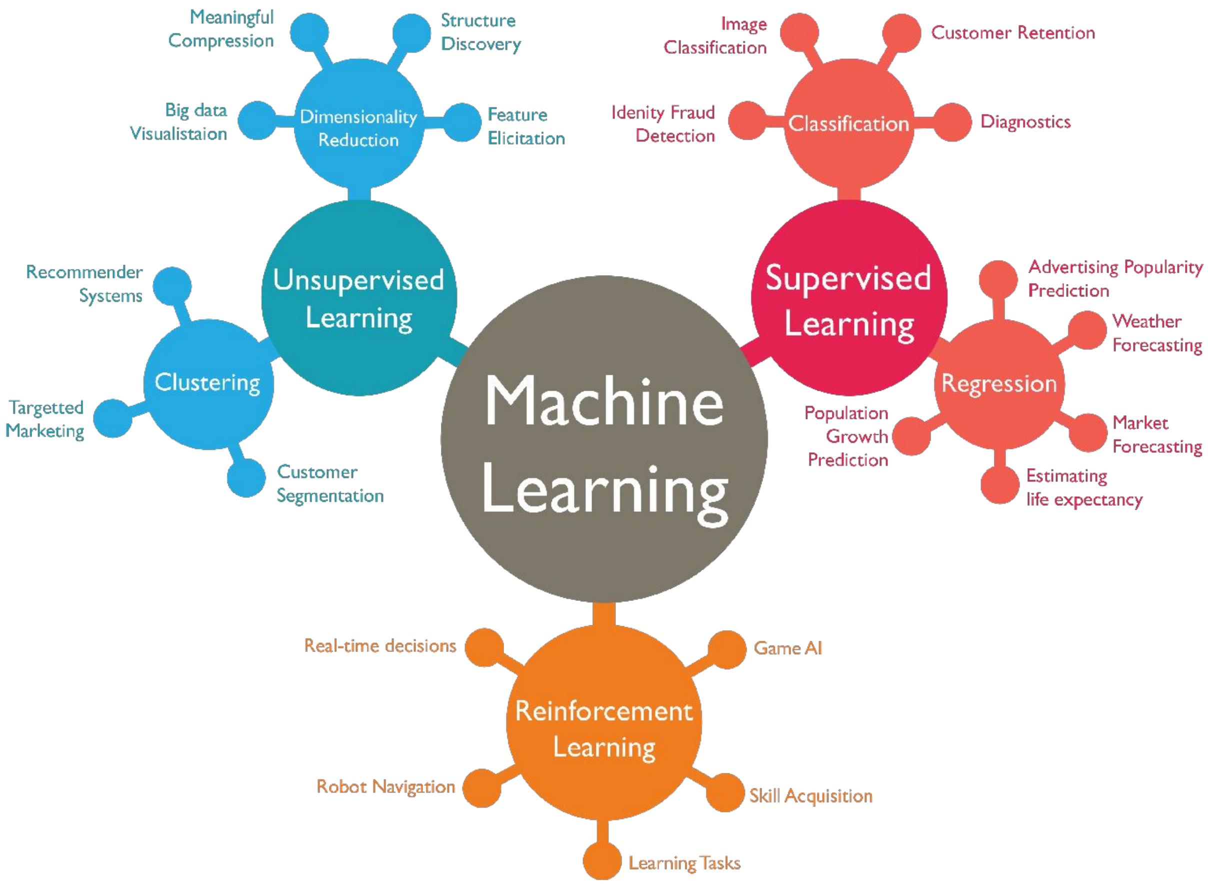
User Case: ML in Health Science

Providing universal access to AI education and practice

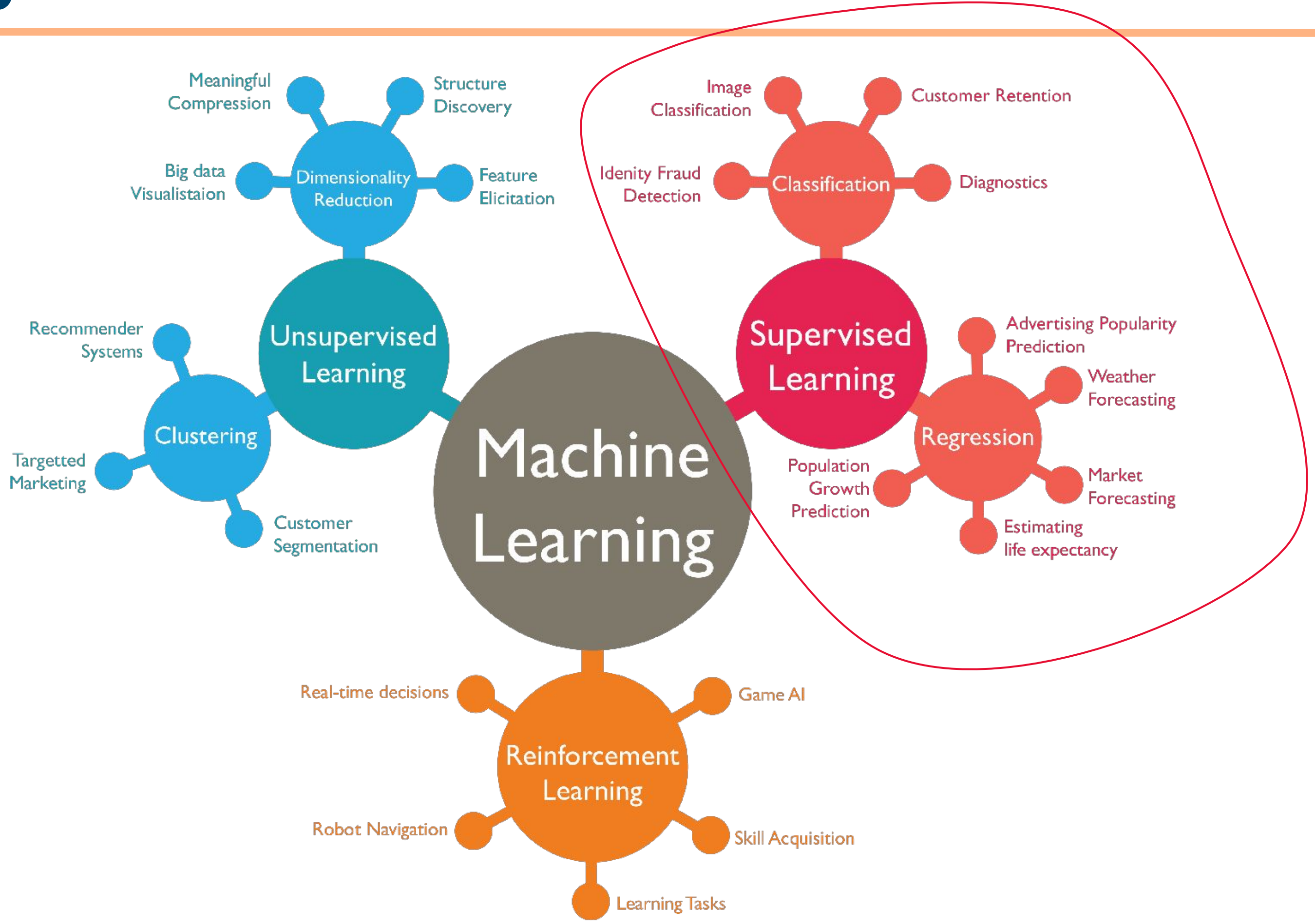
Sum up


Artificial Intelligence
Machine Learning
Data Science
Big Data
Deep Learning

Sum up


Machine Learning
Unsupervised Learning
Dimensionality Reduction
Meaningful Compression
Structure Discovery
Big data Visualistaion
Feature Elicitation
Clustering
Recommender Systems
Targetted Marketing
Customer Segmentation
Supervised Learning
Classification
Image Classification
Customer Retention
Idenity Fraud Detection
Diagnostics
Regression
Advertising Popularity Prediction
Weather Forecasting
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Reinforcement Learning
Real-time decisions
Game AI
Robot Navigation
Skill Acquisition
Learning Tasks

Sum up


Meaningful Compression
Structure Discovery
Big data Visualistaion
Dimensionality Reduction
Feature Elicitation
Unsupervised Learning
Recommender Systems
Clustering
Targetted Marketing
Customer Segmentation
Image Classification
Customer Retention
Idenity Fraud Detection
Classification
Diagnostics
Supervised Learning
Advertising Popularity Prediction
Weather Forecasting
Regression
Population Growth Prediction
Market Forecasting
Estimating life expectancy
Machine Learning
Real-time decisions
Game AI
Reinforcement Learning
Robot Navigation
Skill Acquisition
Learning Tasks

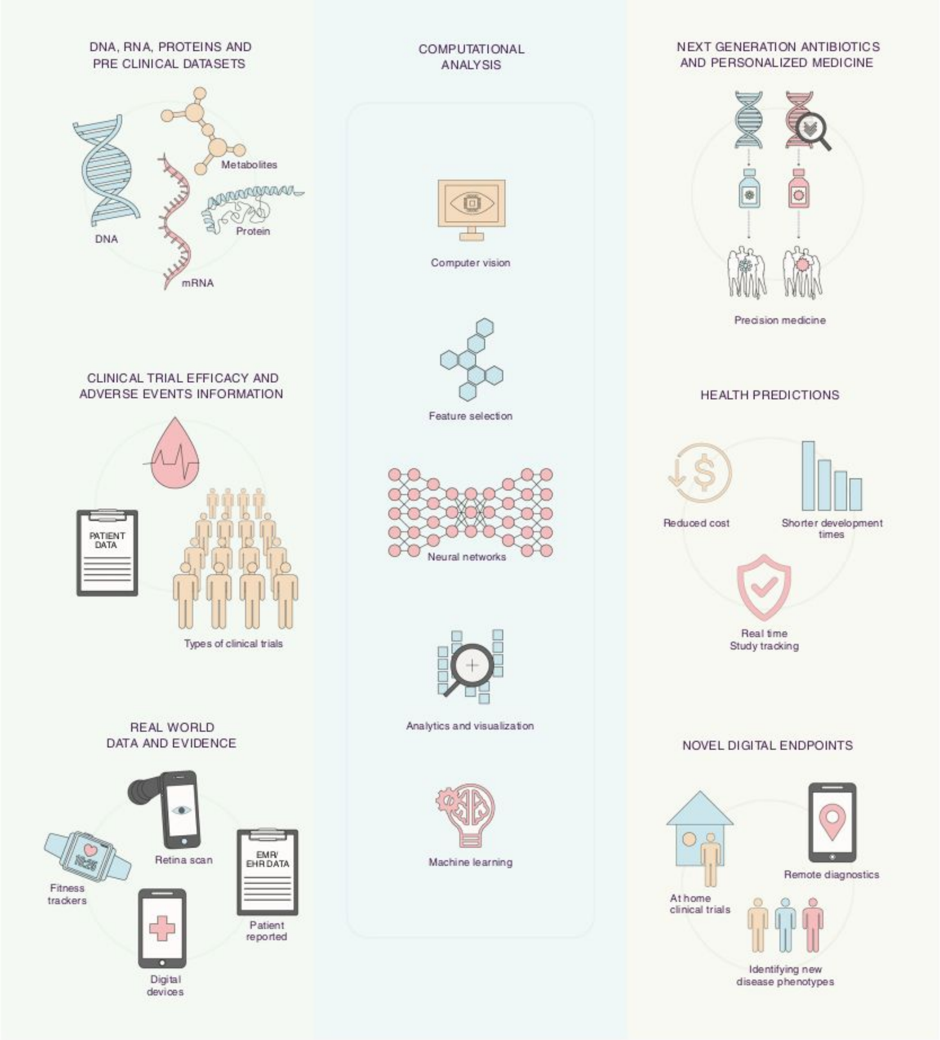


Fig. 1 Use cases of artificial intelligence, computer vision, and machine learning in clinical development

Key areas

- ML based learning to predict pharmaceutical pharmaceutical properties of molecular compounds and targets for drug discovery
- Using pattern recognition and segmentation techniques on medical images (retinal scans, X-ray images..) to enable faster diagnoses and tracking of disease progression
- Developing deep learning techniques on multimodal data sources such as combining genomic and clinical data to detect new predictive models
- Using Natural Language processing techniques to process medical records to tabulate data

Skeleton of a project

DATA PREPROCESSING

ALGORITHM SELECTION

TRAINING

VALIDATION

DATA PREPROCESSING

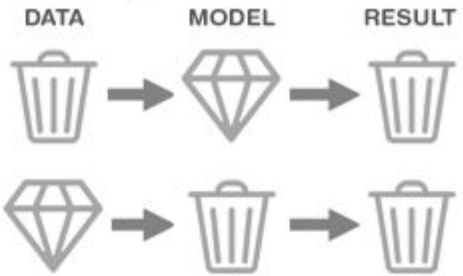
Data preprocessing

DATA CLEANING

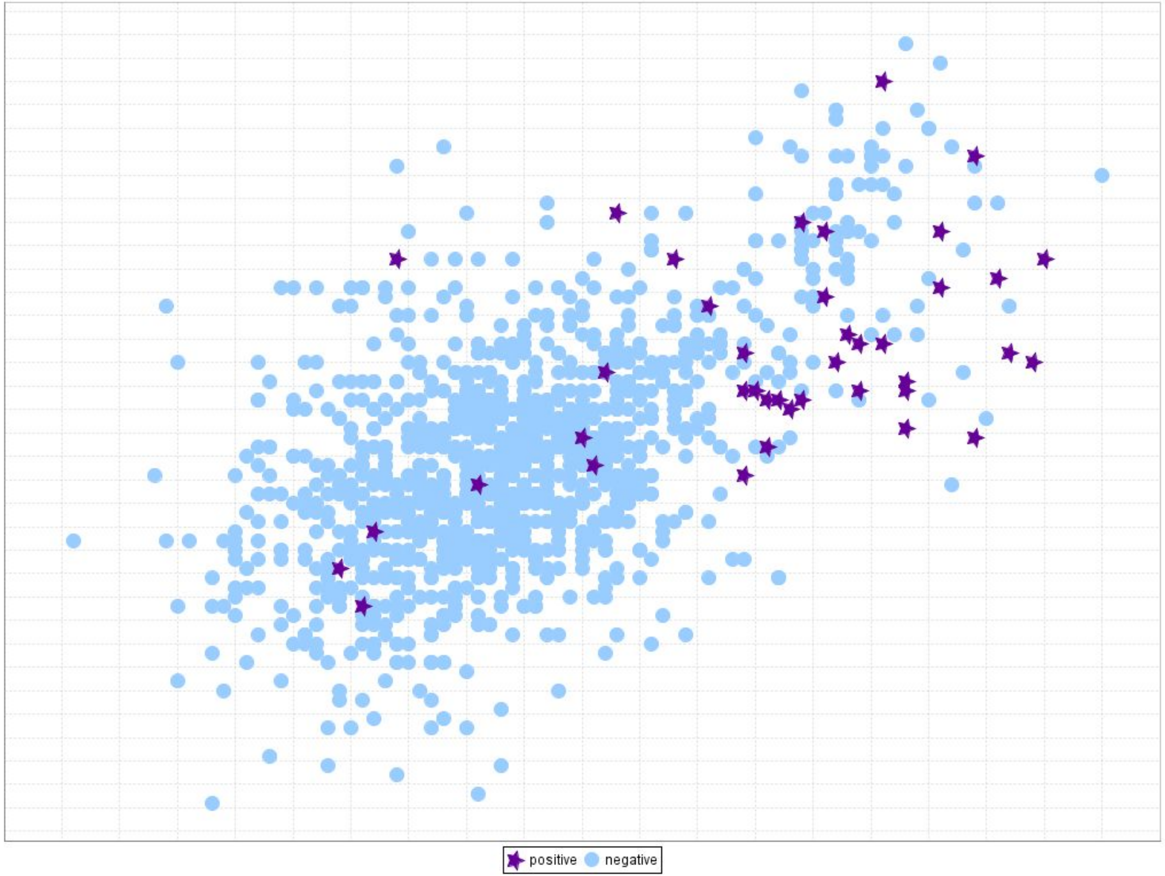
IMBALANCED DATASET

Data cleaning

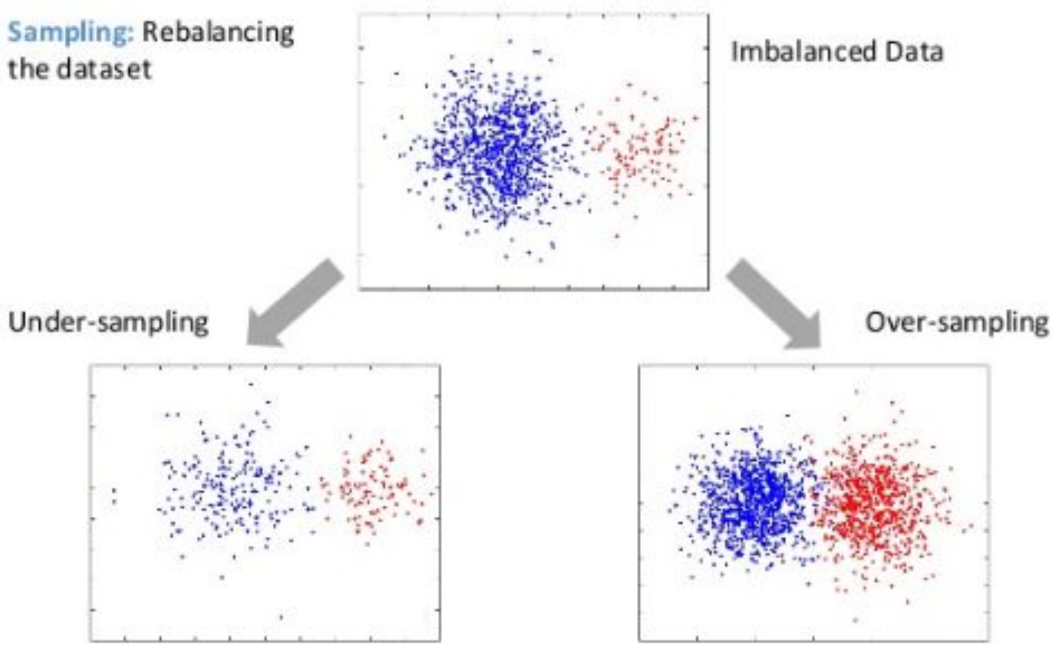
GARBAGE IN, GARBAGE OUT!



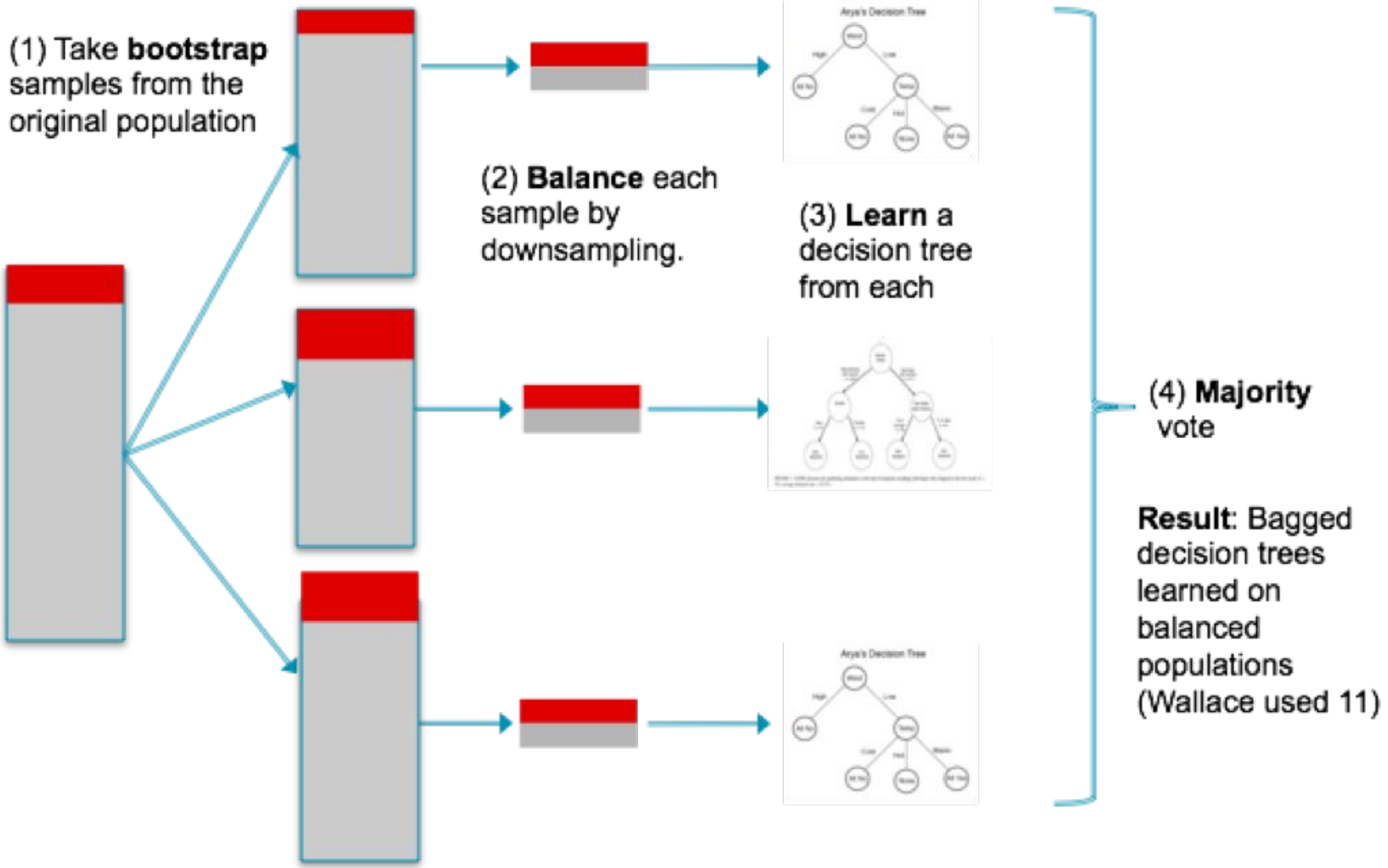
Imbalanced datasets



Imbalanced datasets

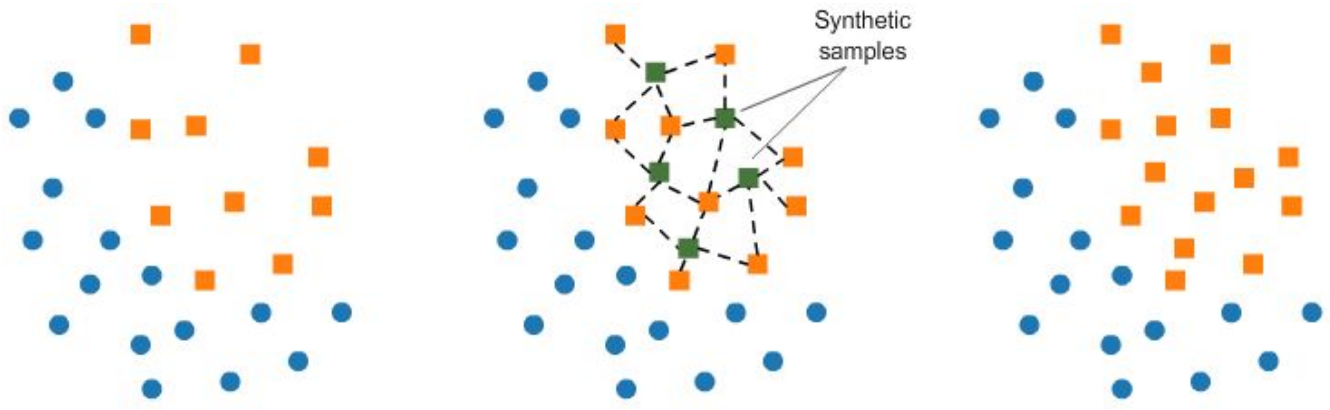


Imbalanced datasets



Imbalanced datasets

SMOTE : https://jair.org/index.php/jair/article/view/10302



ALGORITHM SELECTION

Algorithm selection

Is the tumor malignant or benign ?

Algorithm selection

Blood glucose level at 9am

Algorithm selection

Relation between genetic variant and disease

Algorithm selection

MRI image analysis in lung cancer

Algorithm selection

Is a patient going to have complications if I give treatment?

Algorithm selection

Depends on:

- Type of problem to solve
- Characteristics of the target variable
- ...

VALIDATION OF THE MODEL

Validation of the model

Two elements:

- Metric
- Type of validation

Metrics: classification

Confusion Matrix: a table showing correct predictions and types of incorrect predictions.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Metrics: classification

Confusion Matrix: a table showing correct predictions and types of incorrect predictions.

True Positive:

Interpretation: You predicted positive and it's true.

You predicted that a woman is pregnant and she actually is.

True Negative:

Interpretation: You predicted negative and it's true.

You predicted that a man is not pregnant and he actually is not.

False Positive: (Type 1 Error)

Interpretation: You predicted positive and it's false.

You predicted that a man is pregnant but he actually is not.

False Negative: (Type 2 Error)

Interpretation: You predicted negative and it's false.

You predicted that a woman is not pregnant but she actually is.

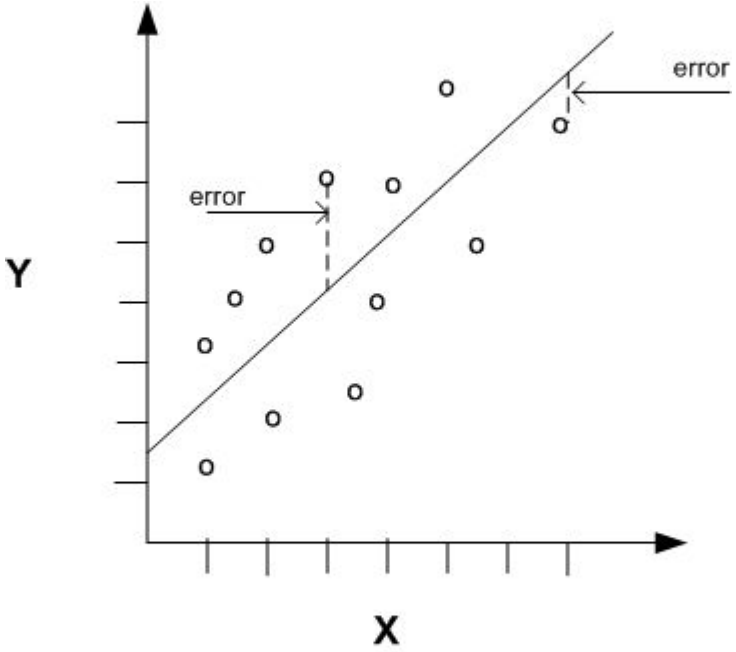
Predicted Values \ Actual Values	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Metrics: classification

Metric	Formula	Evaluation Focus
Accuracy	$\mathrm{ACC} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$	Overall effectiveness of a classifier
Error rate	$\mathrm{ERR} = \frac{\mathrm{FP}+\mathrm{FN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$	Classification error
Precision	$\mathrm{PRC} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}$	Class agreement of the data labels with the positive labels given by the classifier
Sensitivity Specificity	$\mathrm{SNS} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}$ $\mathrm{SPC} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}$	Effectiveness of a classifier to identify positive labels How effectively a classifier identifies negative labels
ROC	$ROC = \frac{\sqrt{SNS^2+SPC^2}}{\sqrt{2}}$	Combined metric based on the Receiver Operating Characteristic (ROC) space [53]
F_1 score	$F_1 = 2\frac{PRC \cdot SNS}{PRC+SNS}$	Combination of precision (PRC) and sensitivity (SNS) in a single metric
Geometric Mean	$GM = \sqrt{SNS \cdot SPC}$	Combination of sensitivity (SNS) and specificity (SPC) in a single metric

Metrics: regression

R squared: is a statistical measure of how close the data are to the fitted regression line.



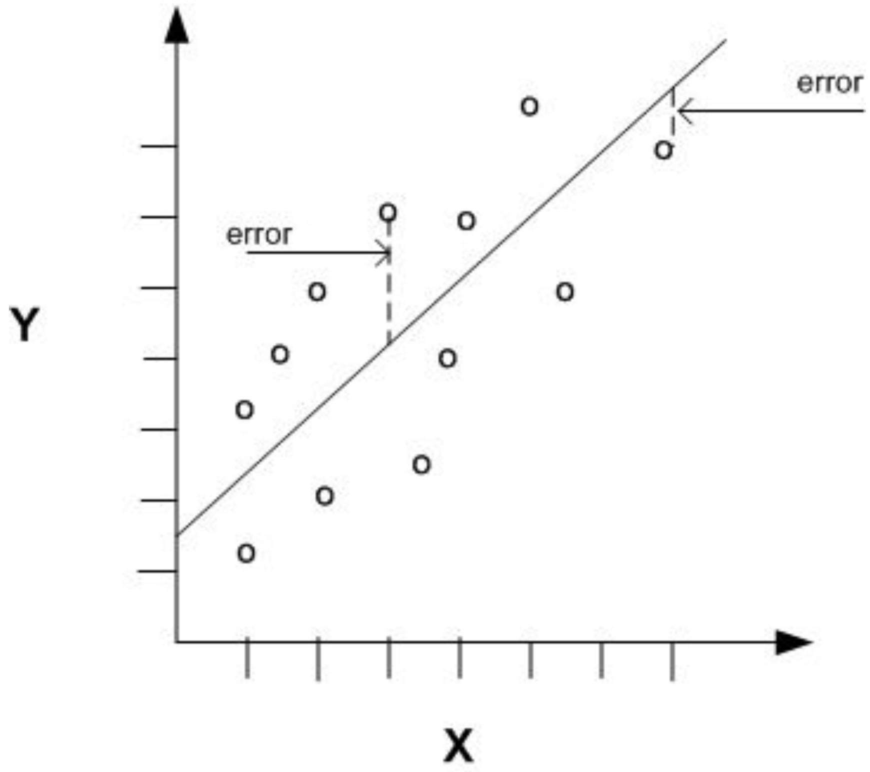
Metrics: regression

R-squared is the percentage of the response variable variation that is explained by a linear model.

Or: R-squared = Explained variation / Total variation

R-squared is always between 0 and 1:

- 0 indicates that the model explains none of the variability of the response data around its mean.
- 1 indicates that the model explains all the variability of the response data around its mean.



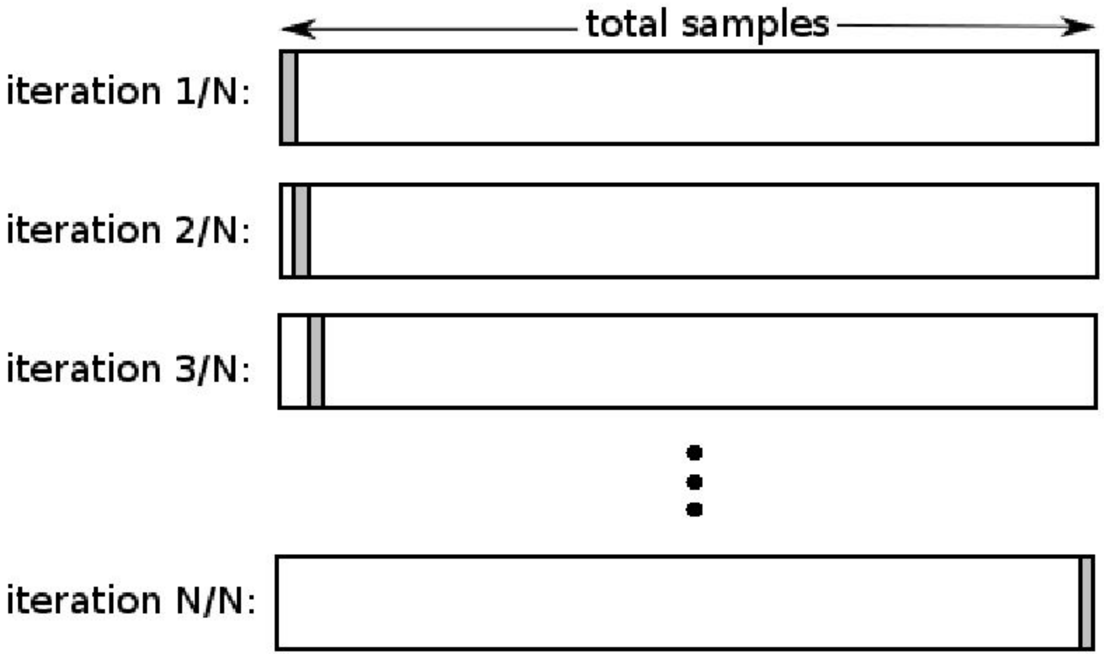
Validation: internal

K- fold Cross Validation

Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

Validation: internal

Leave One Out Cross Validation



Validation: external

Using a control dataset or a dataset with the same characteristics but from other source.

DISCOVERY

REPLICATION

Exercise

1. Working with imbalanced data sets
   Download the dataset and train a logistic regression model. Build the confusion matrix and write the code to calculate all the metrics.
   https://www.kaggle.com/mlg-ulb/creditcardfraud/home?source=post_page---------------------------
   Try to solve the imbalance with the two techniques mentioned.

2. First small project
Following all the steps explained in the lesson build a project using the following breast cancer dataset from sklearn.