CCM

Introduction to Machine Learning and AI Boot Camp

Hands-on tutorials to help build skills required in Precision Child Health, data science and bioinformatics for SickKids staff and trainees.

Topics include:

- Introduction to machine learning
- Coding practice: regression and classification models
- Performance metrics and validation
- Neural network architectures
- Model training and optimization
- Open-source repositories

Prerequisite: Prior knowledge of R is required to follow along with the coding practice section.

To register please RSVP at:
https://ccm20251111.eventbrite.com/

Hands-on Bioinformatics Tutorials for Biologists

Tuesday Nov 11, 2025
10 AM – 1 PM

Multimedia Room, PGCRL 3rd floor, or on Zoom

Organized by the Centre for Computational Medicine (ccm.sickkids.ca) at the SickKids Research Institute with support from other groups:

SickKids Research Training Centre — www.sickkids.ca/research

Digital Research Alliance of Canada — alliancecan.ca

Compute Ontario — computeontario.ca

November 11, 2025

(A Whirlwind) Introduction to Machine Learning in R

Agenda

Introduction to machine learning (ML)

ML approaches

- Supervised learning
- Unsupervised learning

Dataset overview

Coding practice

Introduction to Machine Learning

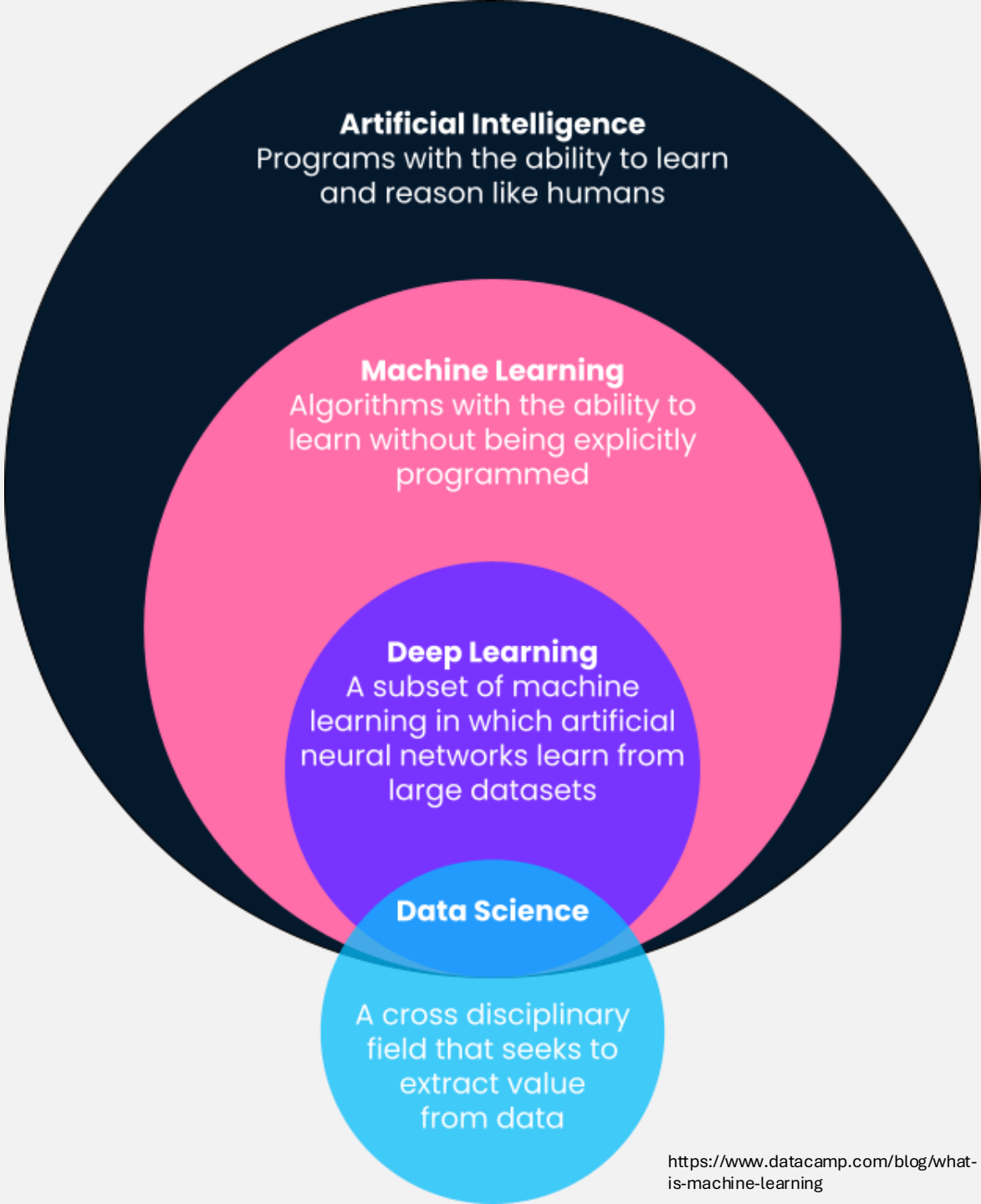
What is machine learning (ML)?

A branch of artificial intelligence that involves the development and evaluation of statistical models and algorithms that learn from data without following explicit instructions.

ML has applications across industries:

- Disease prediction
- Drug discovery
- Financial risk assessment
- Computer vision for self-driving cars
- Supply chain optimization
- Social media personalization

…and many others

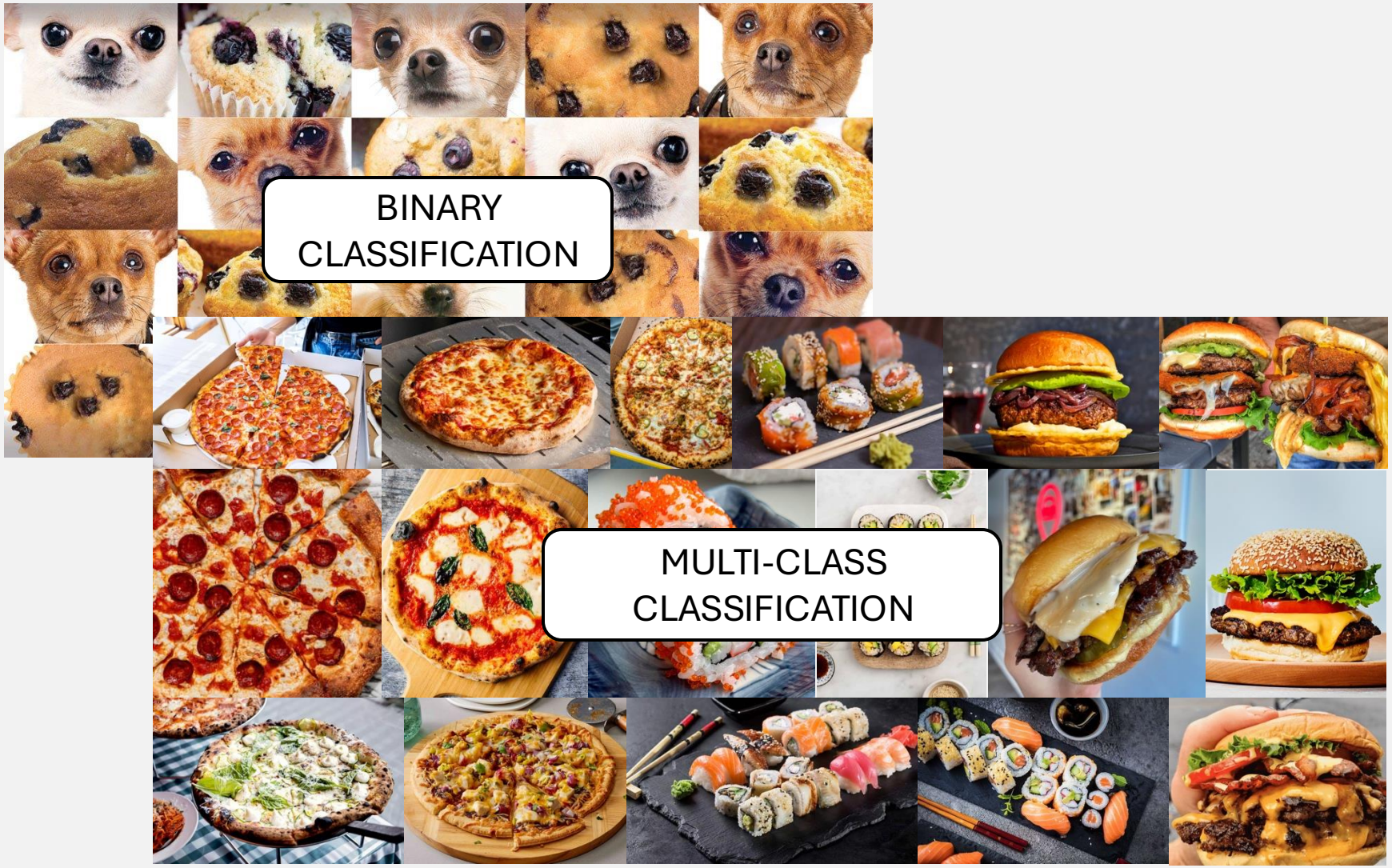


https://www.datacamp.com/blog/what-is-machine-learning

Machine Learning Approaches

Supervised Learning

- Model is trained on labeled data.
- Classification: predict categorical outputs. Can be binary or multi-class.
- Regression: predict numerical values.
- Algorithms include:
  - Regression
  - Support vector machines
  - Decision trees



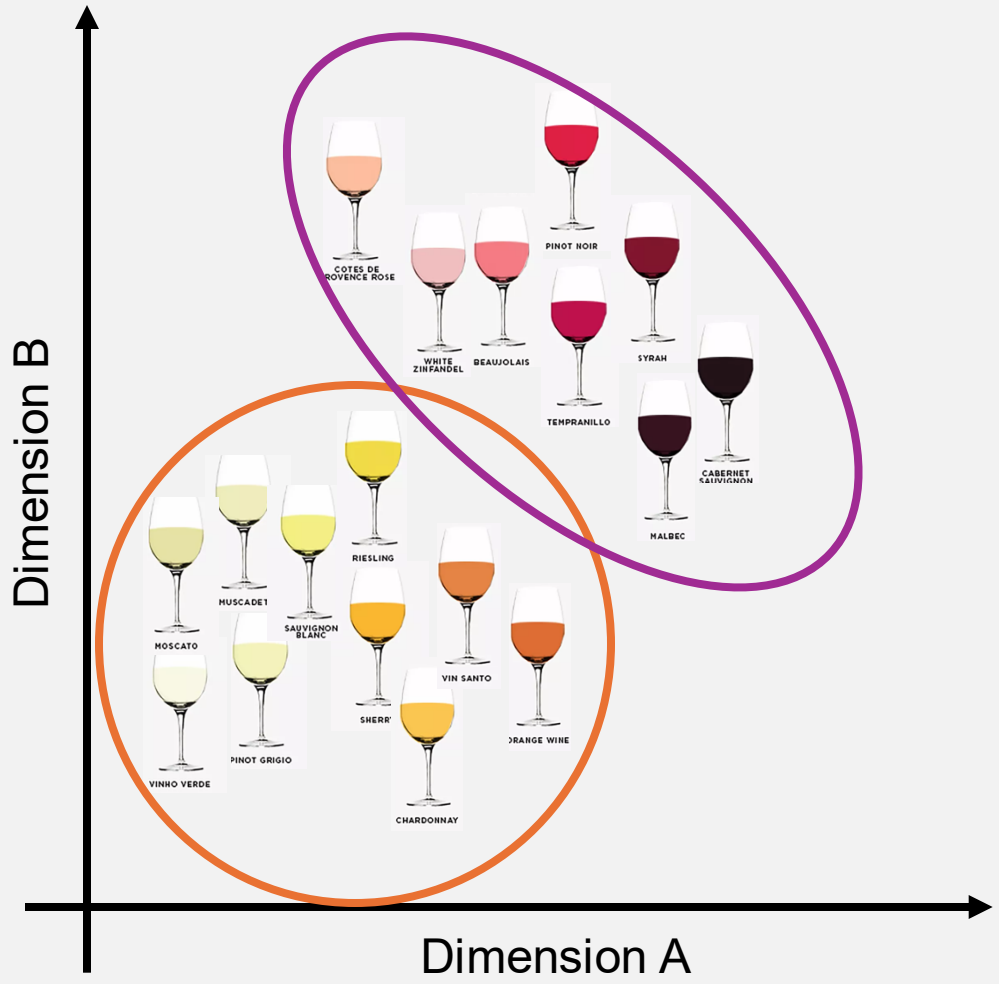
Machine Learning Approaches

Supervised Learning

- Model is trained on labeled data.
- Classification: predict categorical outputs. Can be binary or multi-class.
- Regression: predict numerical values.
- Algorithms include:
  - Regression
  - Support vector machines
  - Decision trees

Unsupervised Learning

- Model is trained on *un*labeled data.
- Clustering: find sub-groups in the data.
- Dimensionality reduction: reduce the number of features in the dataset without losing meaningful information.
- Algorithms include:
  - K-means clustering
  - Hierarchical clustering
  - Principal components analysis



Machine Learning Approaches

Supervised Learning

- Model is trained on labeled data.
- Classification: predict categorical outputs. Can be binary or multi-class.
- Regression: predict numerical values.
- Algorithms include:
  - Regression
  - Support vector machines
  - Decision trees

Unsupervised Learning

- Model is trained on *un*labeled data.
- Clustering: find sub-groups in the data.
- Dimensionality reduction: reduce the number of features in the dataset without losing meaningful information.
- Algorithms include:
  - K-means clustering
  - Hierarchical clustering
  - Principal components analysis

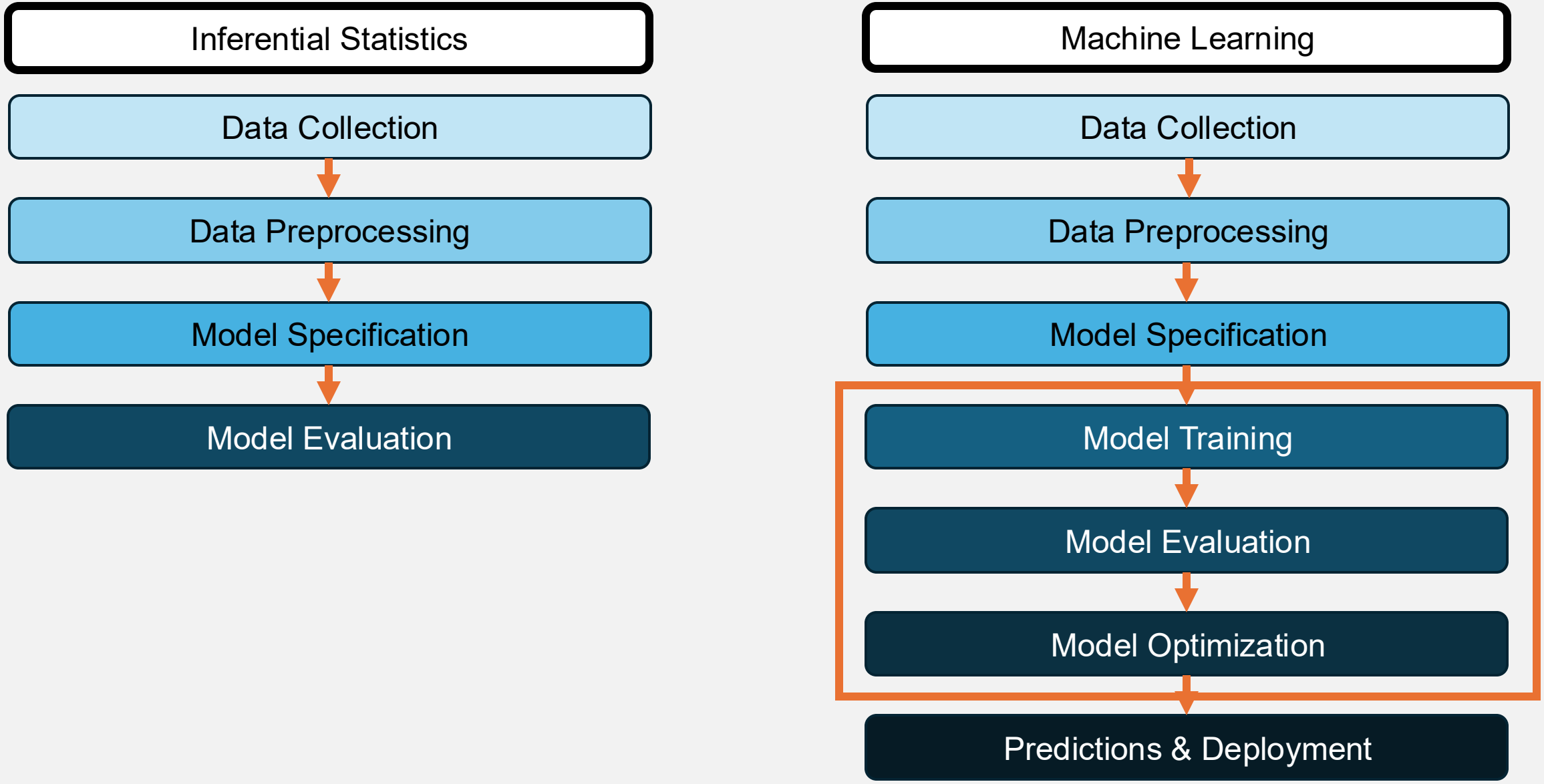
Semi-supervised Learning

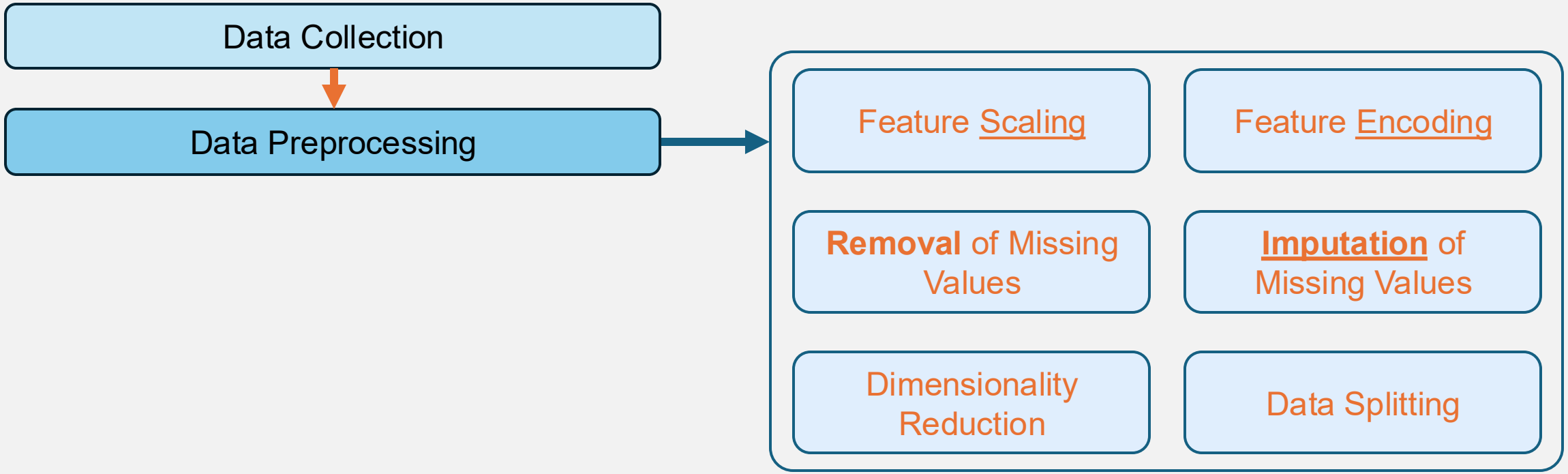
- Model is trained on both labeled and unlabeled data.
- Unlabeled data help identify patterns in the dataset while labeled data establish structure and guide learning (how many classes in the dataset?).
- Useful when working with limited or incompletely labeled data.

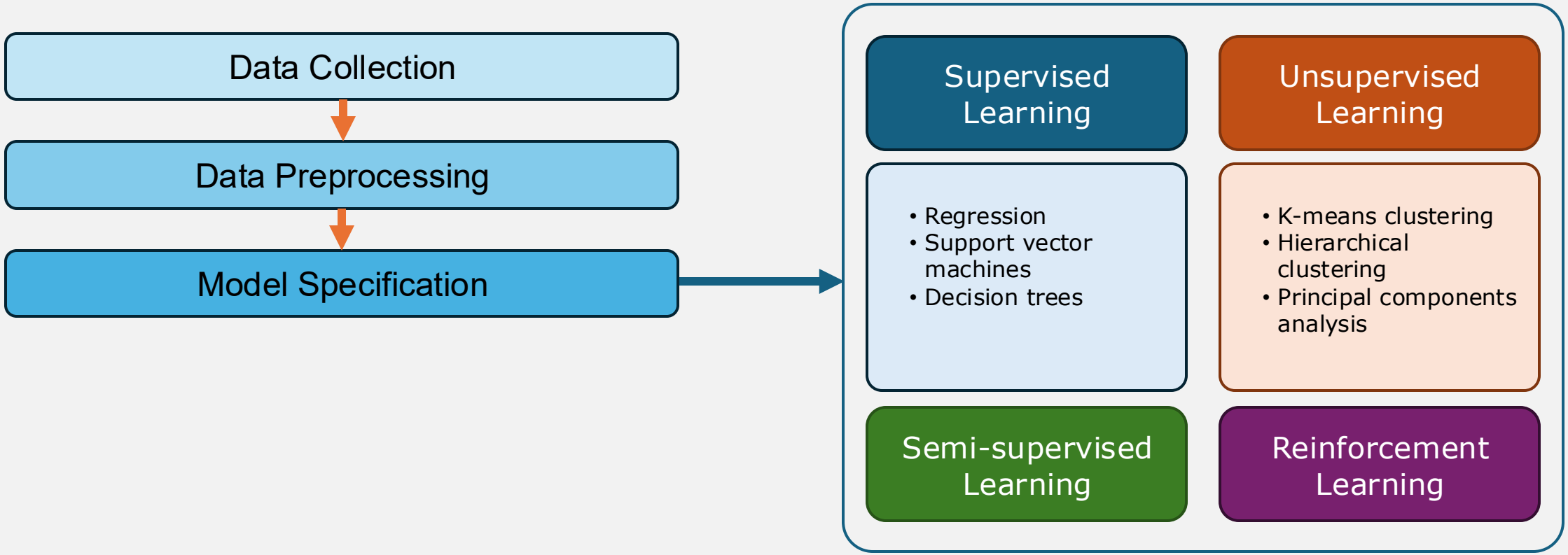
Reinforcement Learning

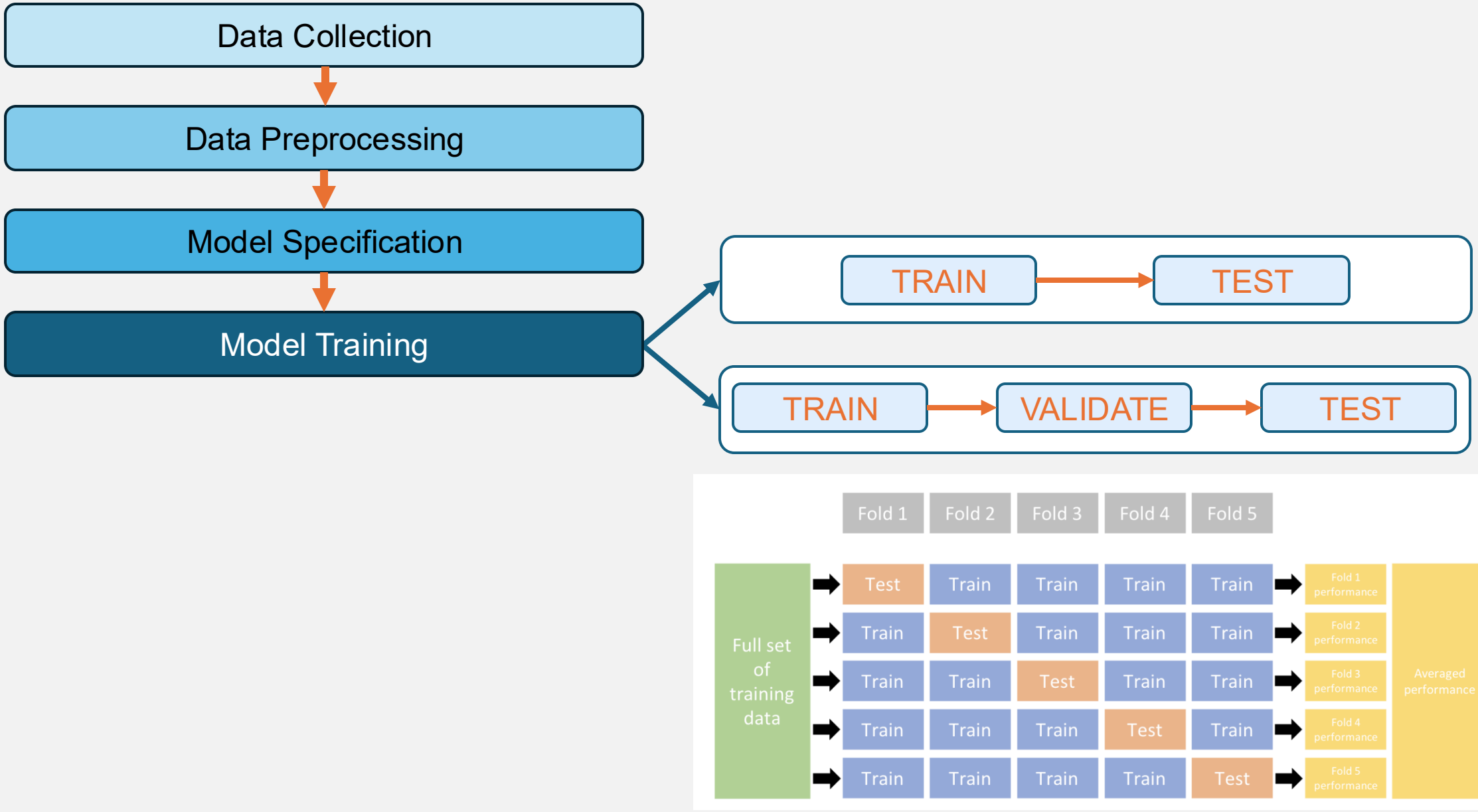
- Model is trained to learn from positive (reward) and negative (punishment) feedback.
- Algorithms include:
  - Value iteration
  - Markov decision process
  - Q learning
- Useful when learning can be sequential.

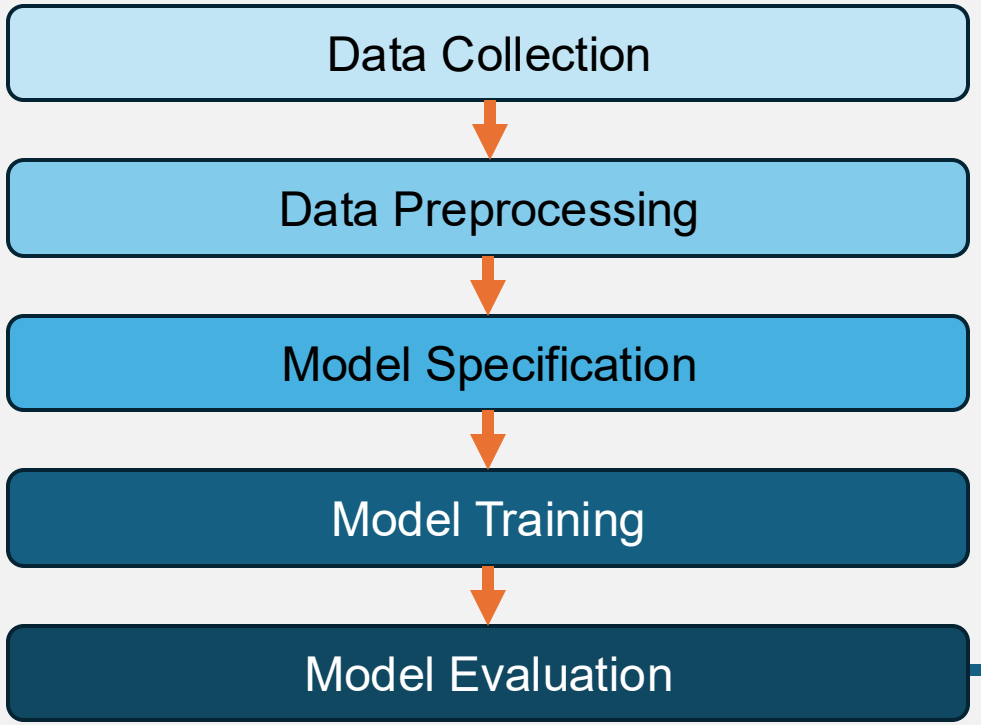
Introduction to Machine Learning




Data Collection
Data Preprocessing
Feature Scaling
Feature Encoding
Removal of Missing Values
Imputation of Missing Values
Dimensionality Reduction
Data Splitting


Data Collection
Data Preprocessing
Model Specification
Supervised Learning
Unsupervised Learning
• Regression
• Support vector machines
• Decision trees
• K-means clustering
• Hierarchical clustering
• Principal components analysis
Semi-supervised Learning
Reinforcement Learning


Data Collection
Data Preprocessing
Model Specification
Model Training
TRAIN
TEST
TRAIN
VALIDATE
TEST
Fold 1
Fold 2
Fold 3
Fold 4
Fold 5
Full set of training data
Test
Train
Fold 1 performance
Fold 2 performance
Fold 3 performance
Fold 4 performance
Fold 5 performance
Averaged performance



Classification Models

- Classification error
- Cross-entropy
- Mean squared error
- Gini index
- Accuracy
- Precision
- Sensitivity (recall)
- Specificity
- Area under the curve (AUC)

Regression Models

- Mean squared error
- Root mean squared error
- Mean absolute error
- Deviance
- R^2
- Root mean squared log error

		True Outcome	
		Yes	No
Predicted Outcome	Yes	True positive	False positive
	No	False negative	True negative

Data Collection

Data Preprocessing

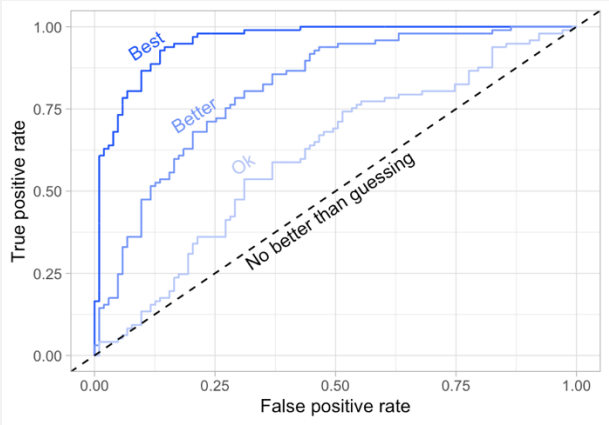
Model Specification

Model Training

Model Evaluation

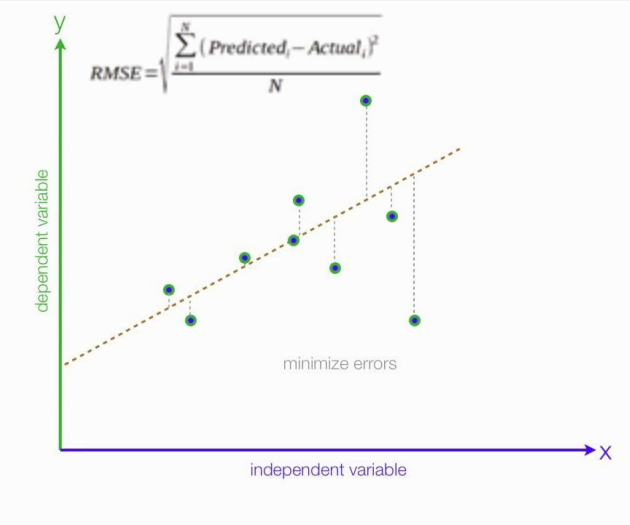
Classification Models

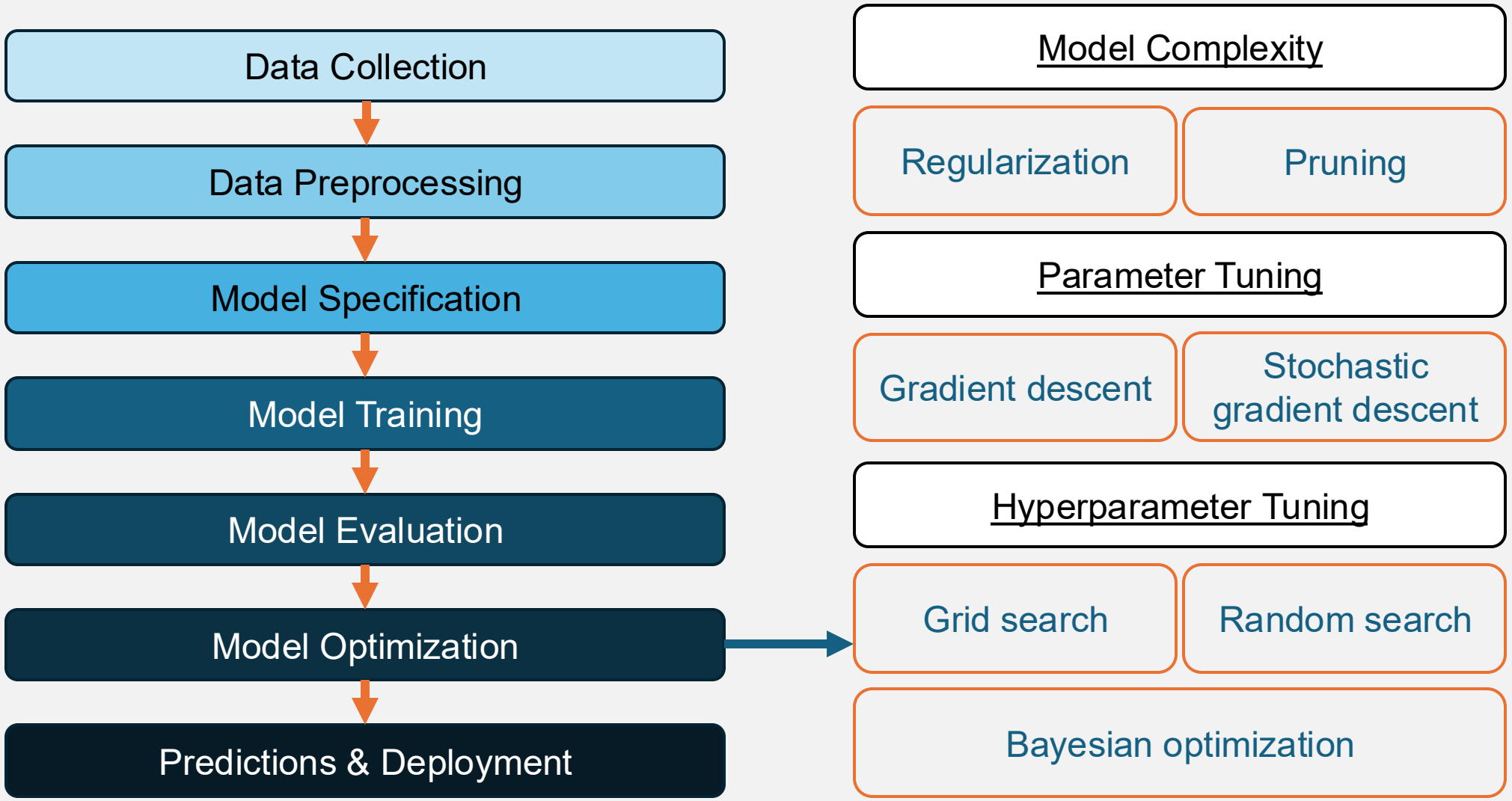
- Classification error
- Cross-entropy
- Mean squared error
- Gini index
- Accuracy
- Precision
- Sensitivity (recall)
- Specificity
- **Area under the curve (AUC)**



Regression Models

- Mean squared error
- **Root mean squared error**
- Mean absolute error
- Deviance
- R^2
- Root mean squared log error




Data Collection
Data Preprocessing
Model Specification
Model Training
Model Evaluation
Model Optimization
Predictions & Deployment
Model Complexity
Regularization
Pruning
Parameter Tuning
Gradient descent
Stochastic gradient descent
Hyperparameter Tuning
Grid search
Random search
Bayesian optimization

CODING PRACTICE

Framingham Heart Study

One of the longest prospective epidemiological studies of cardiovascular disease and its risk factors.

Began in 1948 in Framingham, MA

- Initially enrolled 5209 men and women aged 29-62 years old.
- Followed them over time, with assessments every 2 years.

Examinations included:

- Detailed medical history
- Physical exams
- Lab tests
- Lifestyle and habits
- Noninvasive imaging

Framingham Heart Study Dataset

Subset of the data with 4000+ records and 16 variables.

(https://www.kaggle.com/datasets/captainozlem/framingham-chd-preprocessed-data)

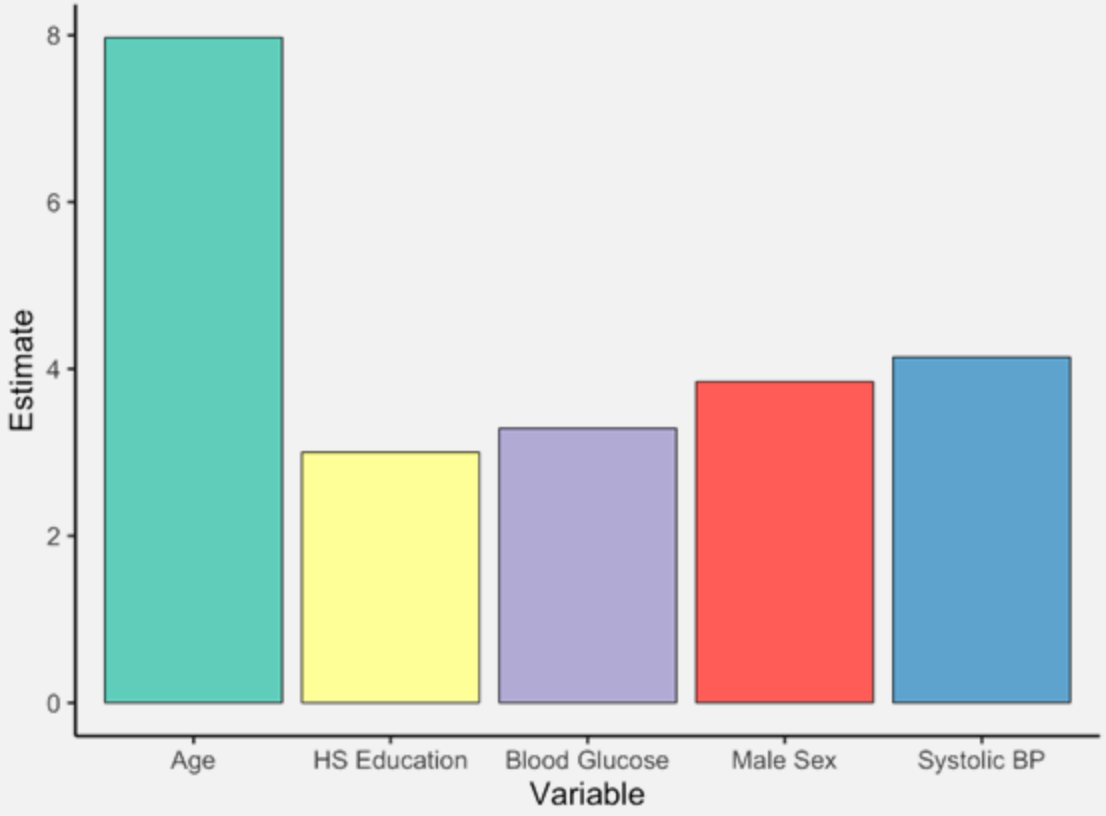
Variable	Description	Class/type
male	Sex (male or female)	Binary
age	Age	Continuous
education	Education (0-11 years, HS or GED, some uni, uni grad+)	Categorical
currentSmoker	Whether or not the patient is currently a smoker	Binary
cigsPerDay	Number of cigarettes smoked per day, on average	Continuous
BPMeds	Whether or not the patient is on blood pressure medication	Binary
prevalentStroke	Whether or not the patient had previously had a stroke	Binary
prevalentHyp	Whether or not the patient is hypertensive	Binary
diabetes	Whether or not the patient has diabetes	Binary
totChol	Total cholesterol level	Continuous
sysBP	Systolic blood pressure	Continuous
diaBP	Diastolic blood pressure	Continuous
BMI	Body mass index	Continuous
heartRate	Heart rate	Continuous
glucose	Glucose level	Continuous
TenYearCHD	10-year risk of coronary heart disease	Binary

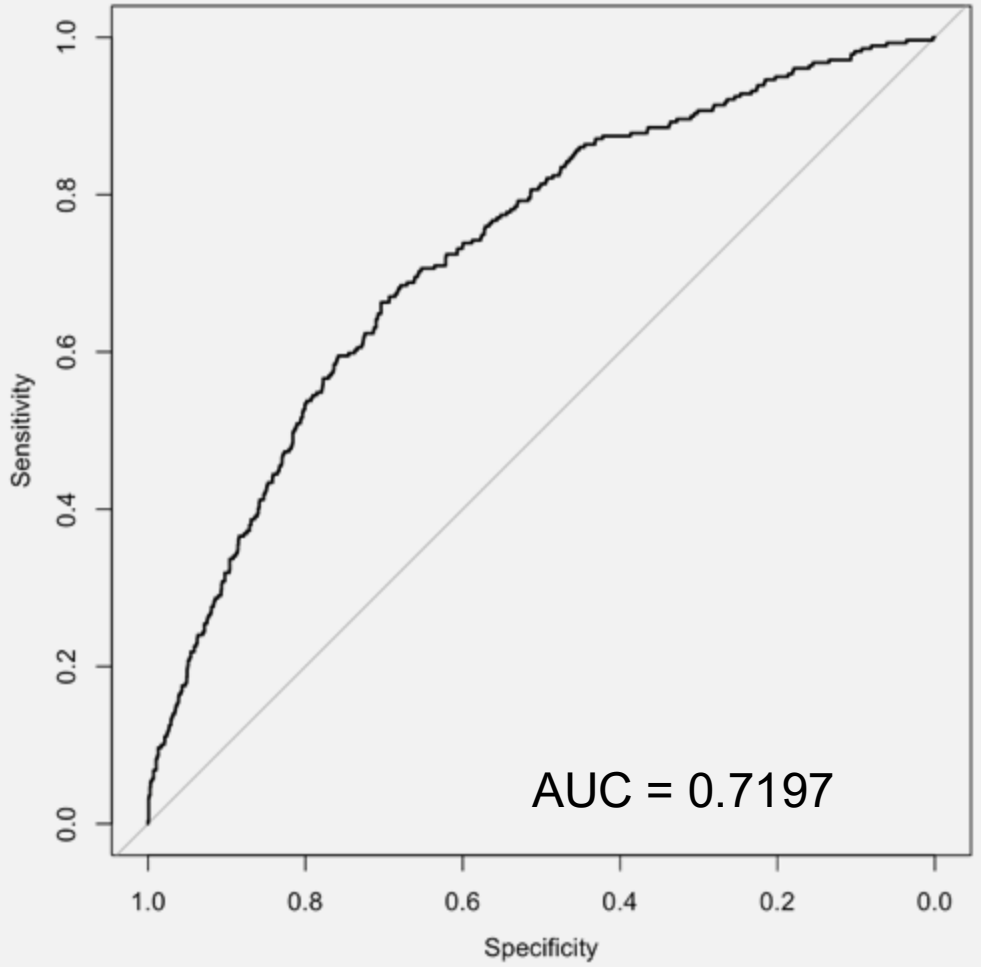
Binary Classification

Logistic Regression

Predictors: all variables

Outcome variable: heart disease



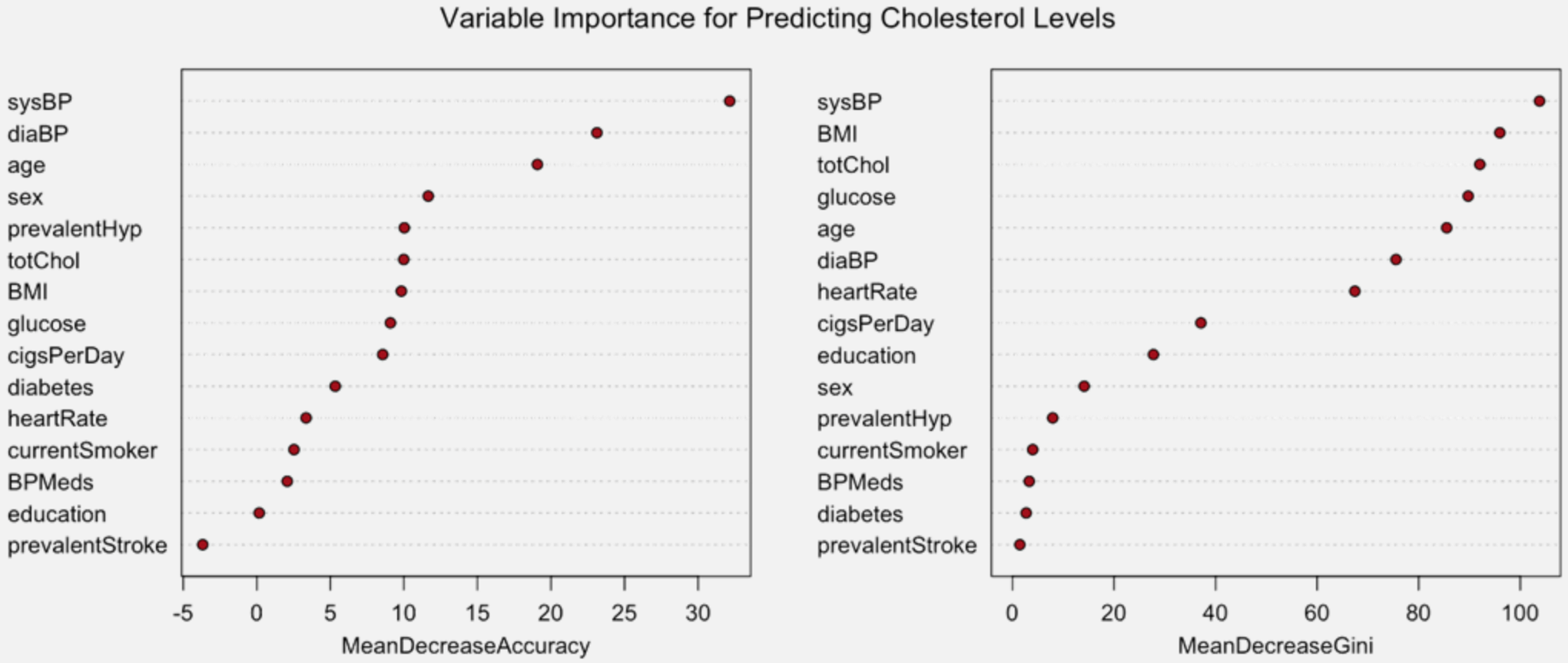


Binary Classification

Random Forest

Predictors: all variables

Outcome variable: heart disease

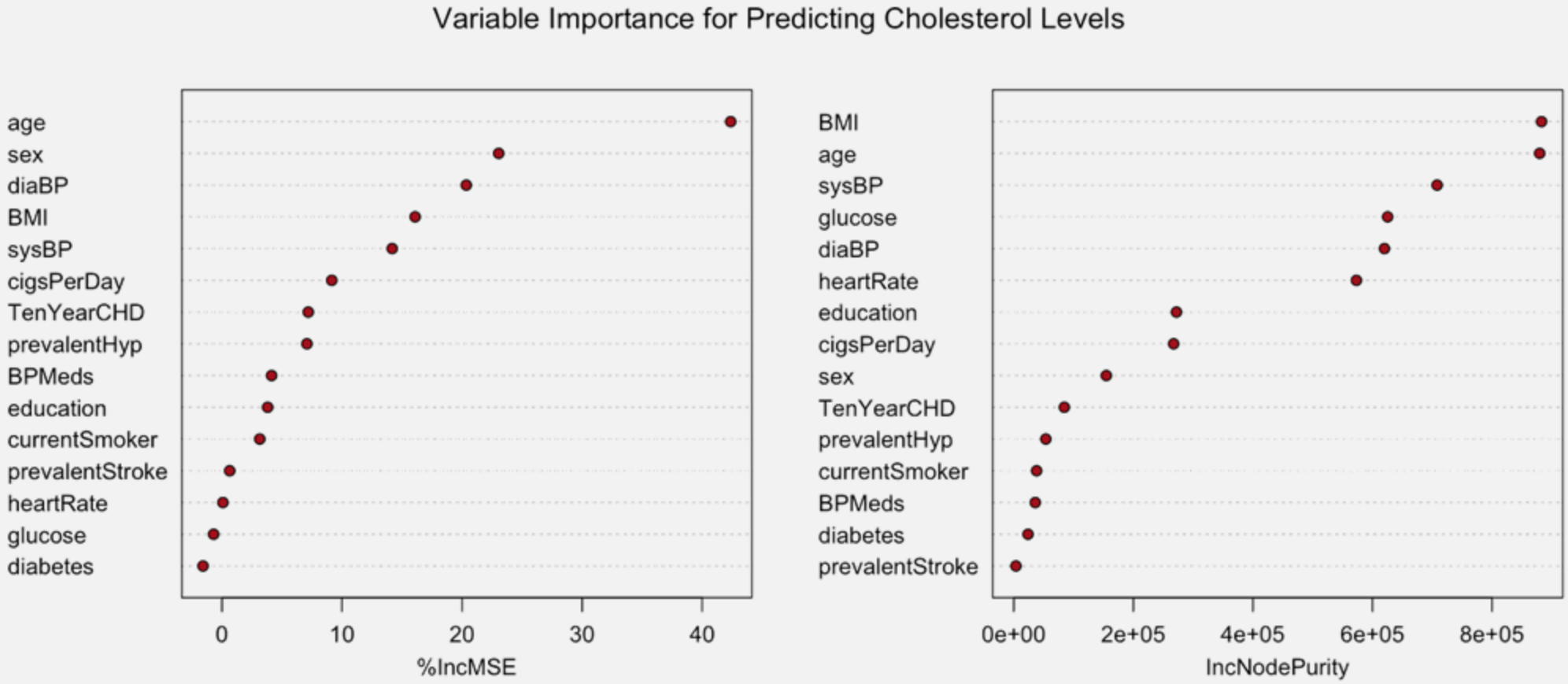


Regression

Random Forest

Predictors: all variables

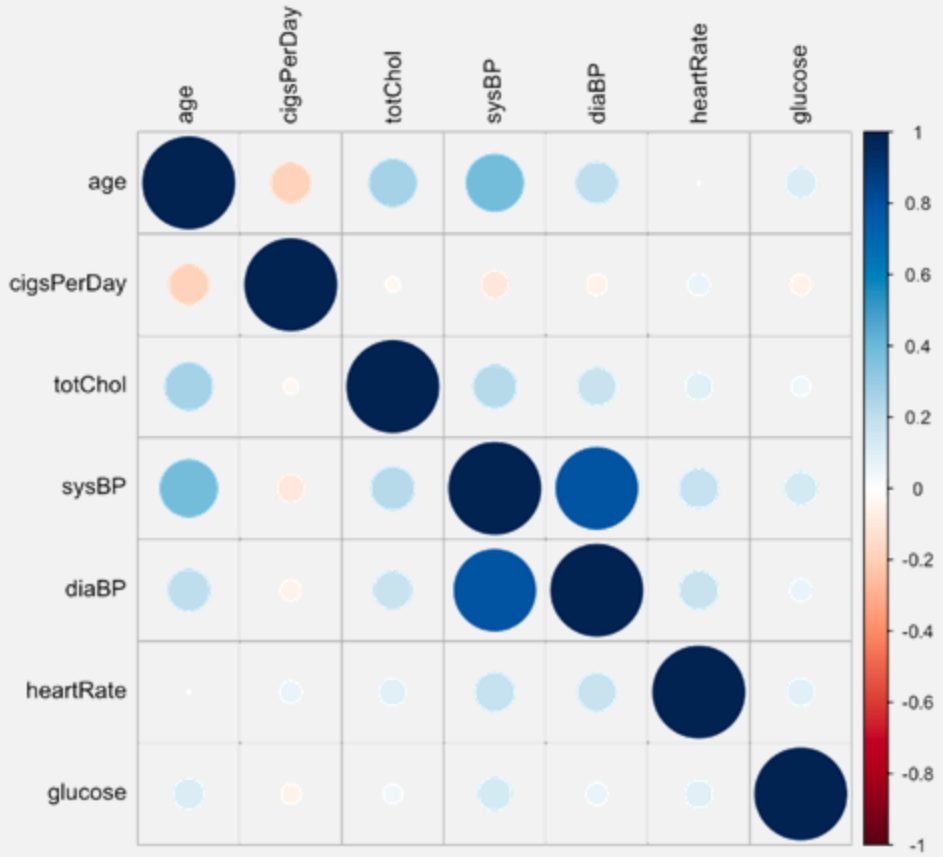
Outcome variable: cholesterol levels

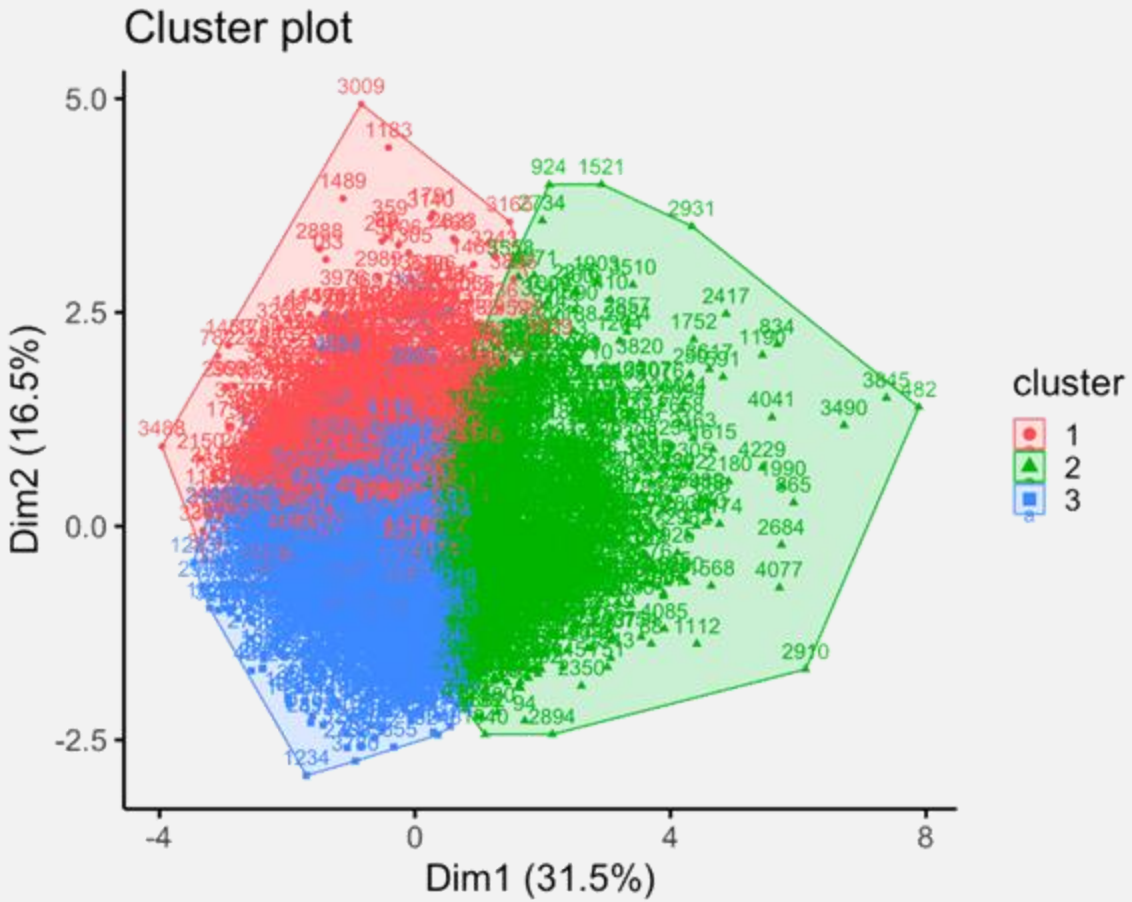


Clustering

K-means

Predictors: all numeric/continuous variables

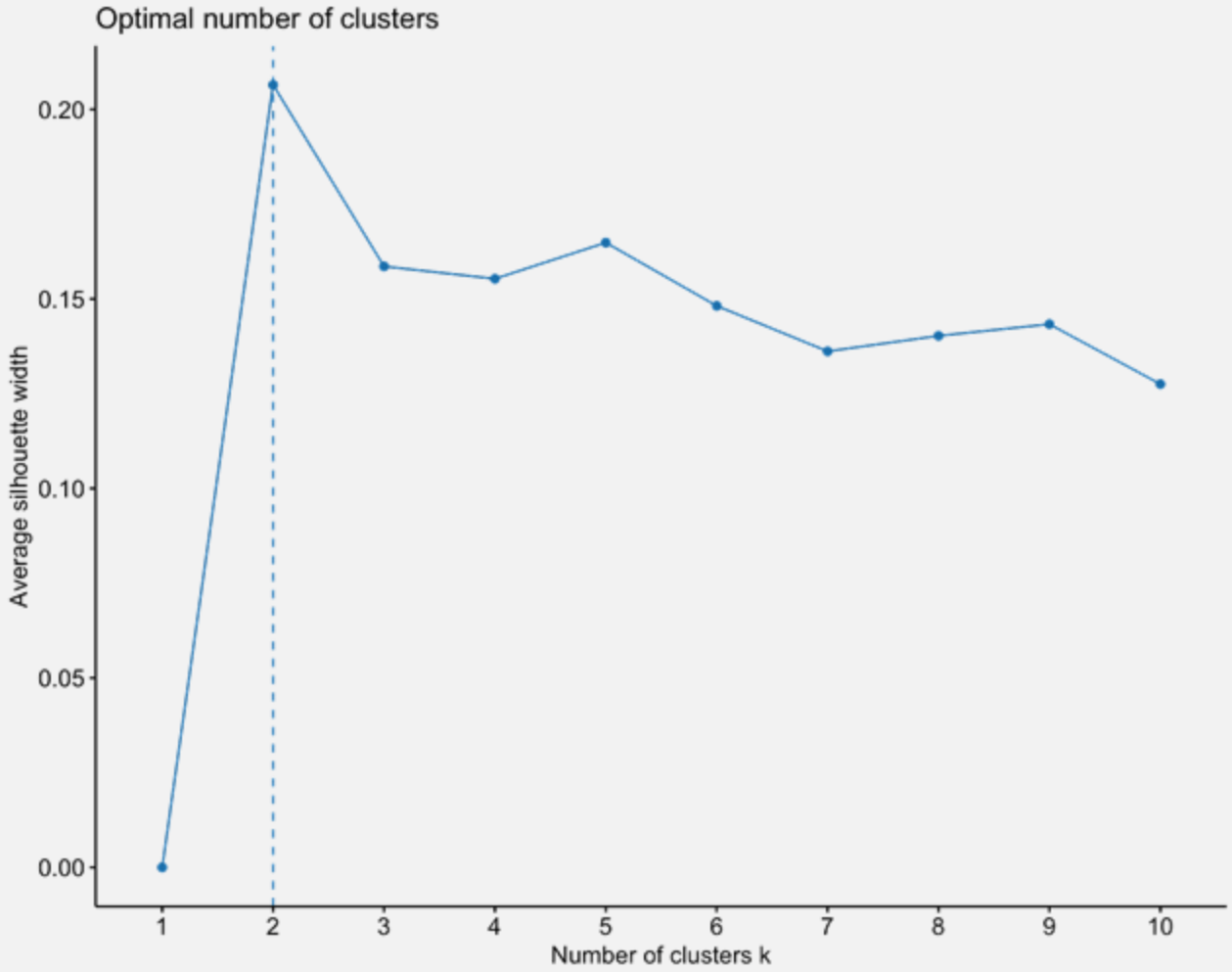


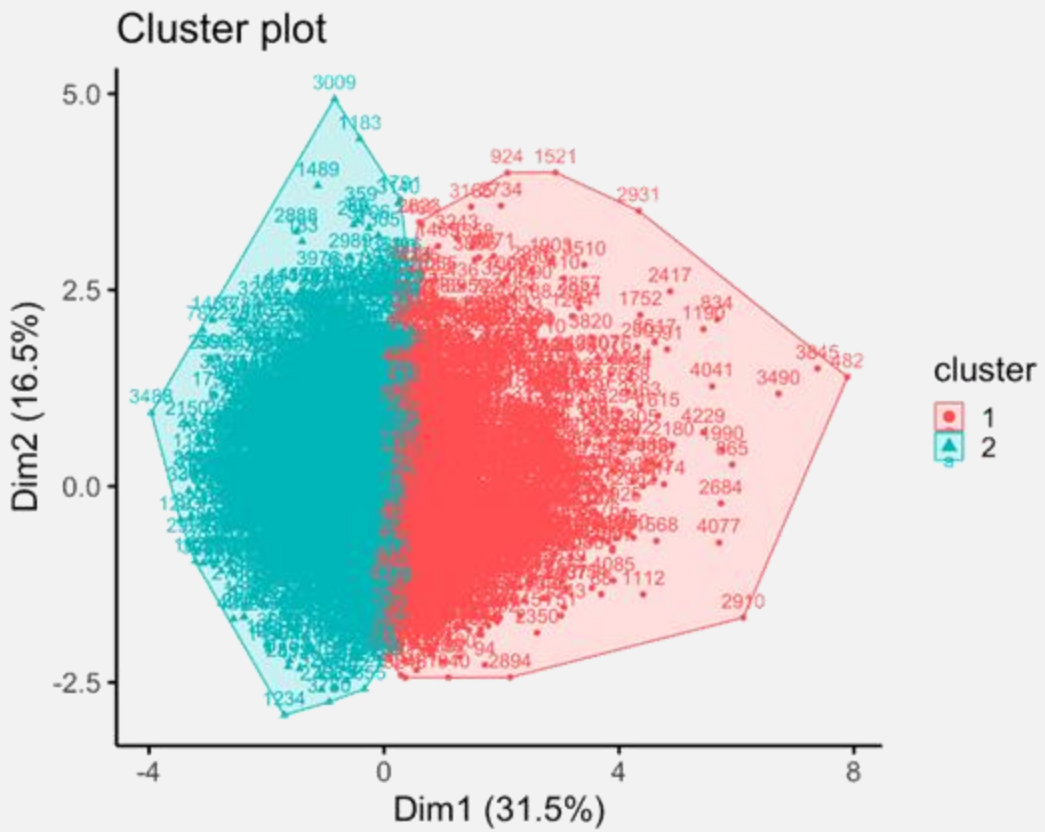


Clustering

K-means

Predictors: all numeric/continuous variables





	Cluster	age	cigsPerDay	totChol	sysBP	diaBP	heartRate	glucose
	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	1	55.2	6.04	258.	151.	91.8	78.7	86.8
2	2	45.7	11.0	223.	120.	76.9	73.7	78.5