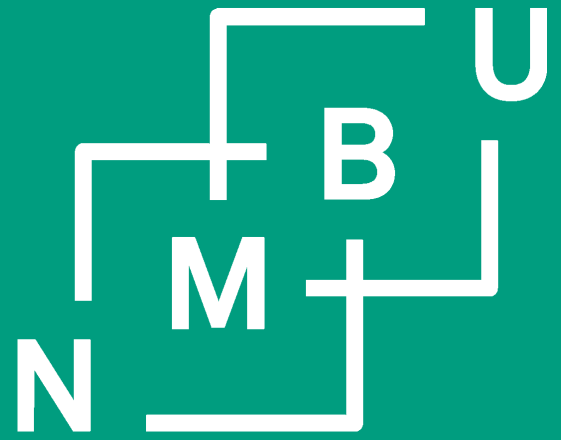
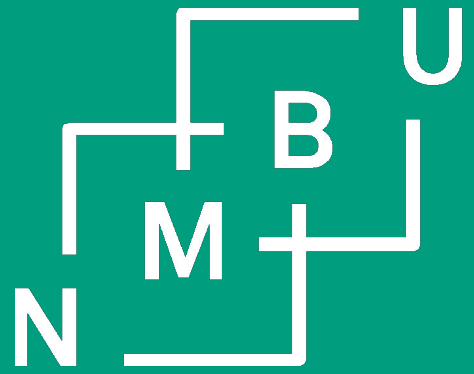

U
B
M
N



Norwegian University
of Life Sciences

Data Handling with Pandas

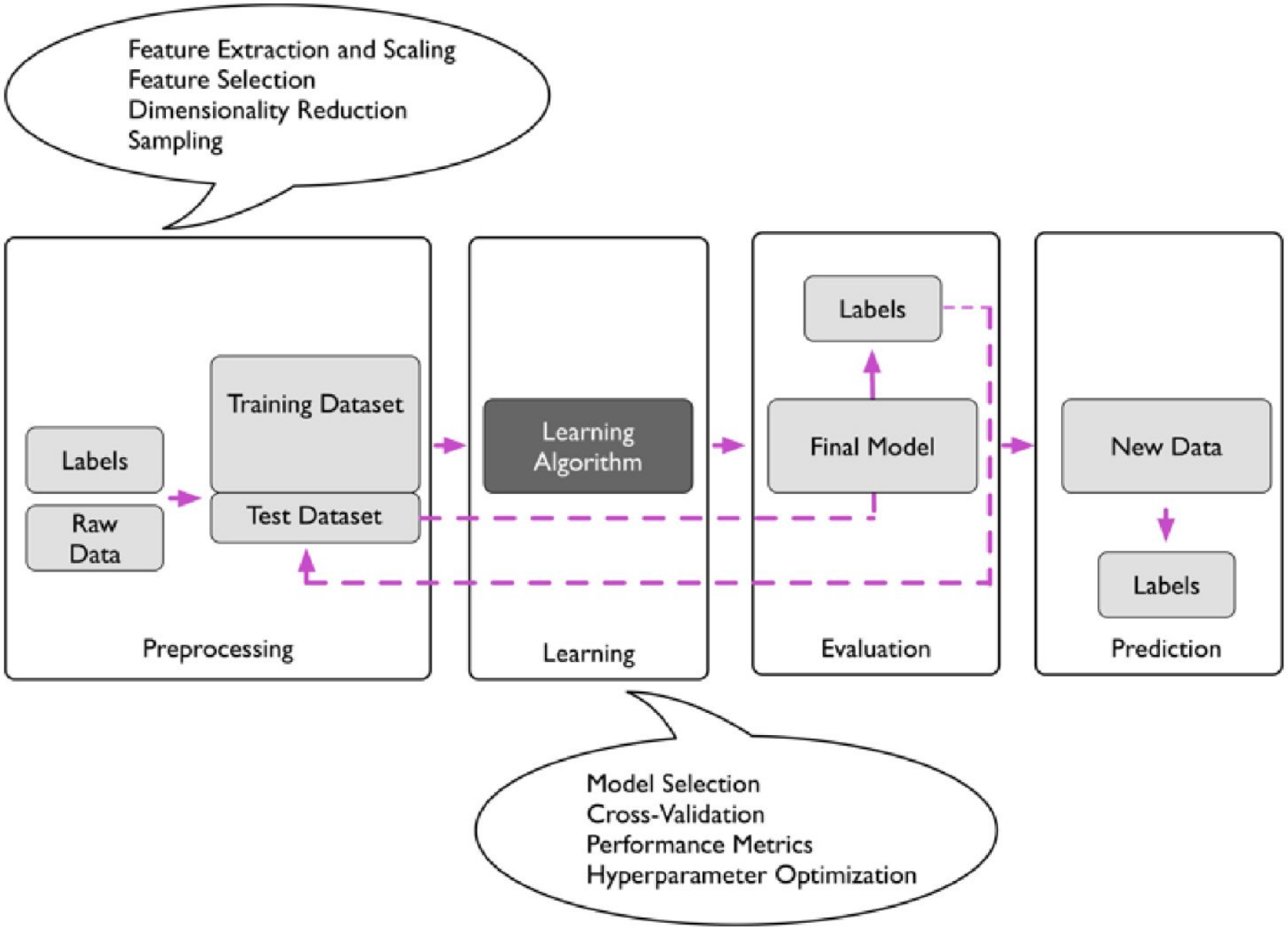
DAT200 - Applied Machine Learning

Department of Data Science, Faculty of Science of Technology

Lecture Agenda

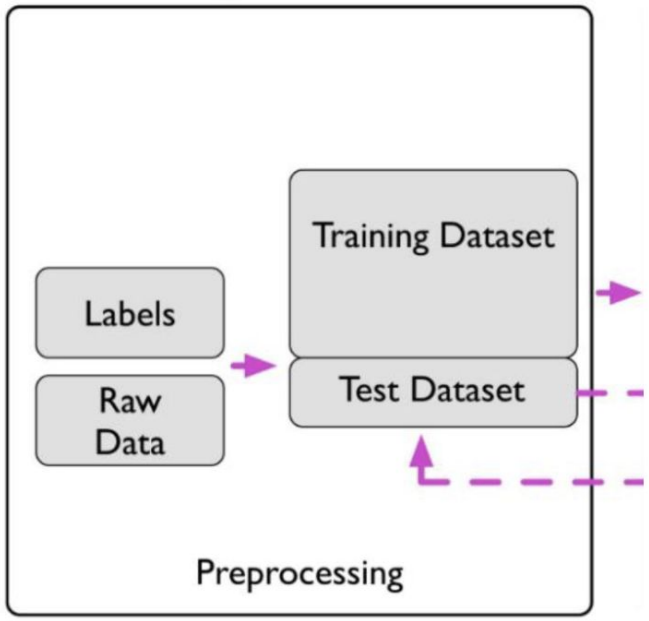
- ML Pipeline
- Prerequisites
- What does the pandas library offer?
- Resources
- Lecture exercises
  - 1
  - 2
  - 3
  - 4
- Compulsory Assignment 1

ML Pipeline


Feature Extraction and Scaling
Feature Selection
Dimensionality Reduction
Sampling
Labels
Raw Data
Training Dataset
Test Dataset
Learning Algorithm
Final Model
Labels
New Data
Labels
Preprocessing
Learning
Evaluation
Prediction
Model Selection
Cross-Validation
Performance Metrics
Hyperparameter Optimization

ML Pipeline: Preprocessing

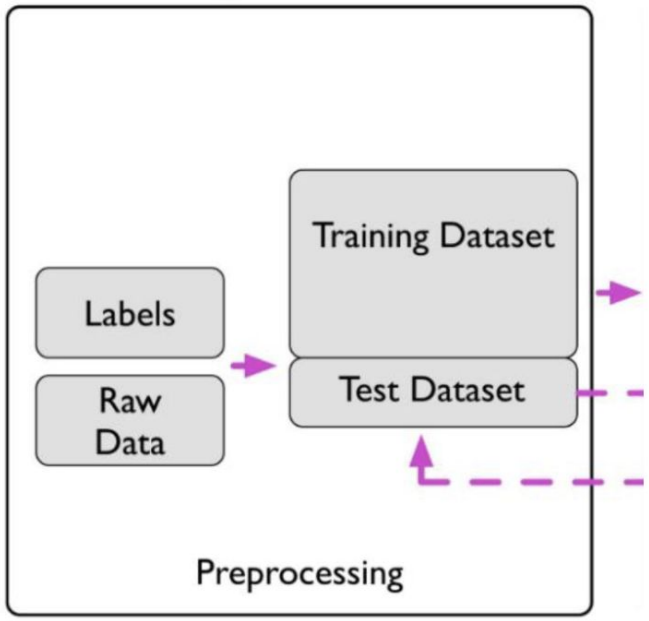
Feature extraction and scaling
Feature selection
Dimensionality reduction
Sampling



- Preprocessing data is one of the most crucial steps in every ML application
- Raw data often needs processing to turn into a good format
- Many ML algorithm require scaling for good performance
- Dimensionality reduction

ML Pipeline: Preprocessing

Feature extraction and scaling
Feature selection
Dimensionality reduction
Sampling

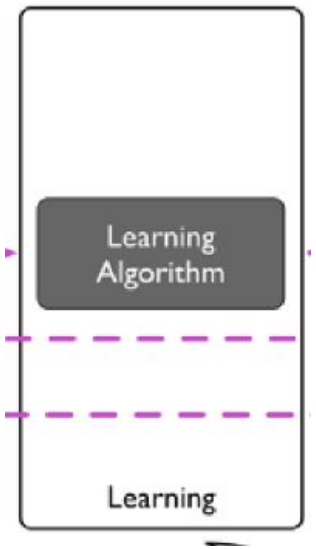


- Separation of data into training and test set
- We want models that generalize well
  - good performance and training AND test set

- Feature engineering
  - Create new features from raw data by transformations

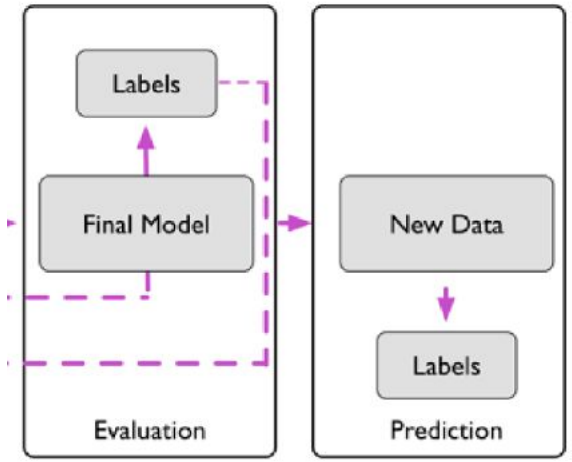
ML Pipeline: Learning algorithm

Training algorithm
Model selection
Cross-validation
Performance metrics
Hyperparameter optimization



- Many available algorithm (ML method, optimization method)
- Each algorithm has strength and weaknesses
- Need to choose performance metric (accuracy, AUC of ROC, etc.)
- Compare performance to find best model
- Cross-validation to test generalization
- Hyperparameter optimization for fine tuning models

ML Pipeline: Evaluation and prediction



- Estimate generalization error with unseen test data
- Track prediction performance with future data

- All transformations applied to the training data are applied to the test data (using the same parameters; parameters can be user set or acquired from an ML algorithm)

Prerequisites

- Assume that you have Anaconda, Miniconda, or Python with a pip package manager installed on your computer.
- Assume that you have installed the required packages
  - The environment YAML file is on Canvas.
    - `conda env create -f dat200_environment.yml`
    - `conda activate dat200_env`
    - `python -m ipykernel install --user --name=dat200_env`

What does the pandas library offer?

- A free and open-source software library for Python
- **Fast** and **highly flexible** structures for handling relational tables, and time series.
- I like to think of it as a tool for handling spreadsheets in Python
- Two primary data structures:
  - Series: 1D array
  - DataFrame (DF): 2D array which can have named rows and columns
- Note that every row/column in a DF is itself a Series
- The library is built on top of Numpy

Resources

- Pandas website and documentation
- Pandas community tutorials (Official pandas website incl. videos)
- RealPython
- Pandas is a powerful tool with many commands and options
  - ChatGPT can be a great tool if you forget the syntax
  - Formulating a precise question of what you want to do is a learning tool in itself

Common tasks with Pandas 1

- Creating a DataFrame
- Indexing the rows and columns
- The big advantage of Pandas is vectorized operations
- **Let us look at some examples**

Lecture exercises 1

- Load Iris dataset into a pandas dataframe from the web
  - https://archive.ics.uci.edu/ml/machine-learning-databases/iris/iris.data
- Set column names to: `'sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'types'`
- Set row names to: flower_1, flower_2, flower_3, …, flower_150

Common tasks with Pandas 2

- Find Unique Values In Pandas Dataframes
- Grouping rows in Pandas
- Create a Column Based on a Conditional in Pandas

Lecture exercises 2

- Find unique values for column `types` in your dataframe
- Compute the column mean for each type
- Create a new column in your dataframe named `sepal_width >= 3` that contains `True` or `False`, depending on whether value in column sepal with is >= 3.0 (`True`) or < 3 (`False`)
- Count how many times sepal width is >= 3 (you can use column `sepal_width >= 3` for that)

Common tasks with Pandas 3

- Filter Pandas Dataframes
- Descriptive Statistics For Pandas Dataframe
- Count values in Pandas Dataframe
- Search A Pandas Column For A Value

Lecture exercises 3

- Count how many times each class occurs (Answer: 50 of each class)
- Create three data subsets from original dataframe (one for setosa, one for versicolor, one for virginica). Use conditional row selection based on column `types`.
- View last 10 rows of columns `sepal_length` and `types`.

Common tasks with Pandas 4

- Dropping Rows And Columns In Pandas Dataframe
- Selecting Pandas DataFrame Rows Based On Conditions
- Sorting Rows In Pandas Dataframes
- Applying Operations Over Pandas Dataframes
- Pivot Tables In Pandas
- Selecting Pandas DataFrame Rows Based On Conditions

Lecture exercises 4

- View rows where `sepal_length` > 5 and `petal_width` < 0.2.
- Make a new DataFrame containing only rows where `petal_width` is exactly 1.8.
- Get descriptive statistics for the whole dataframe and afterward only for column `petal_length`.
- Remove rows named `flower_55` and `flower_77`.
- Remove column `sepal_width >= 3`.
- View all rows of `sepal_length` where `petal_width` is exactly 1.8.
- Get values of the dataframe stored in a numpy array (in practice get rid of columns and rows).
- Remove column `types` and apply a function named `computation` to each cell in dataframe. Function computation should do the following: take the value of the cell, add 1 and multiply that by 3.

Solutions to lecture exercises will be posted to Canvas.

Compulsory Assignment 1

- Solutions to lecture exercises will be posted to Canvas.
- Compulsory Assignment 1 should also be posted to Canvas now.
  - The tasks in CA1 will be similar to the exercises in this lecture
  - Please start with the CAs early, or you can be overwhelmed if you begin right before the deadline.

Thank you for coming!