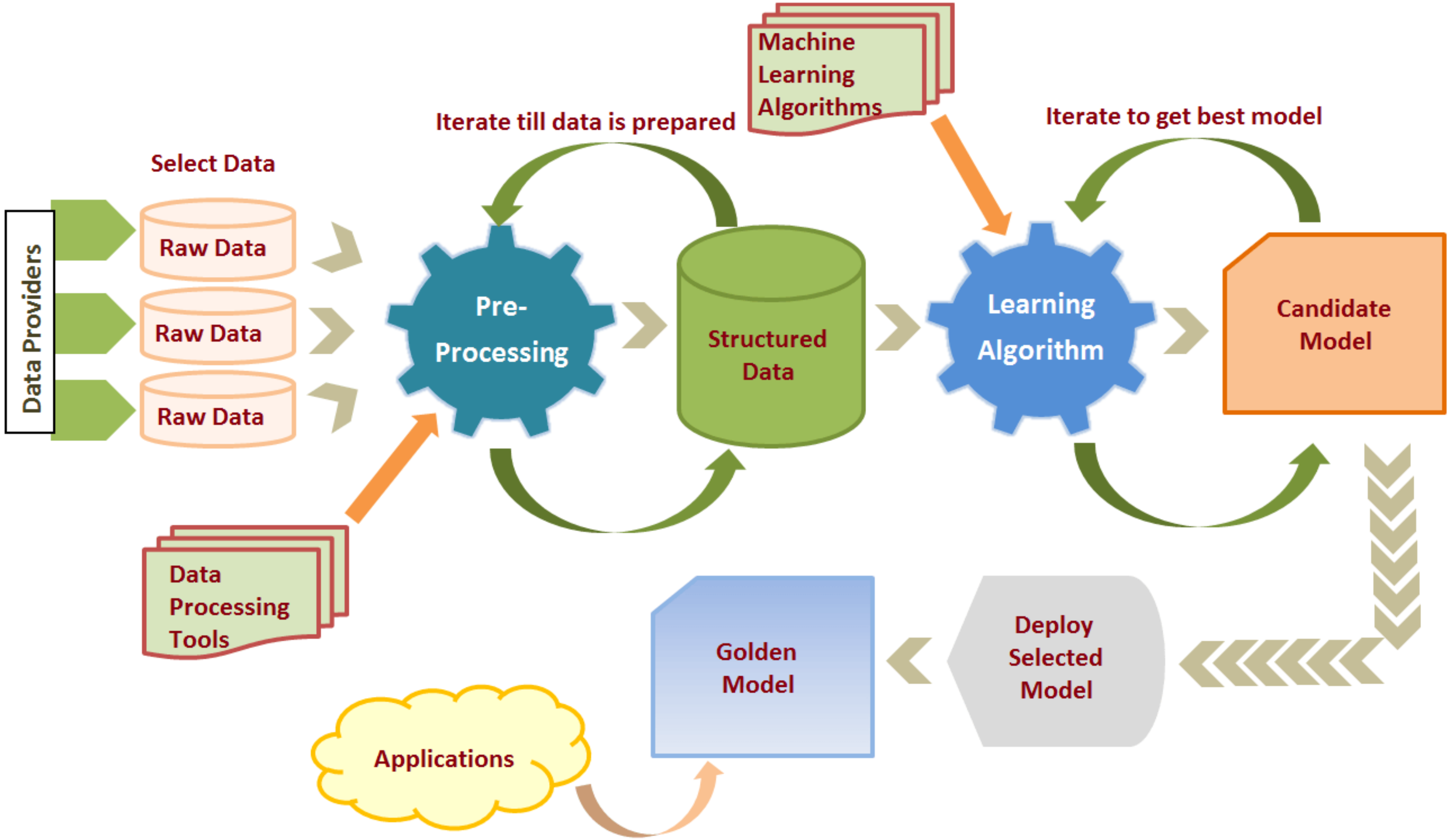
Unit II: Feature Management

CTE309 Machine Learning

AS2024: BE Information Technology

Machine Learning Pipeline



Overview

- Introduction
- Define feature
- Data types in ML
- Feature engineering
- Data scrubbing vs Feature engineering and their relationship
- Importance of Feature Engineering

Introduction

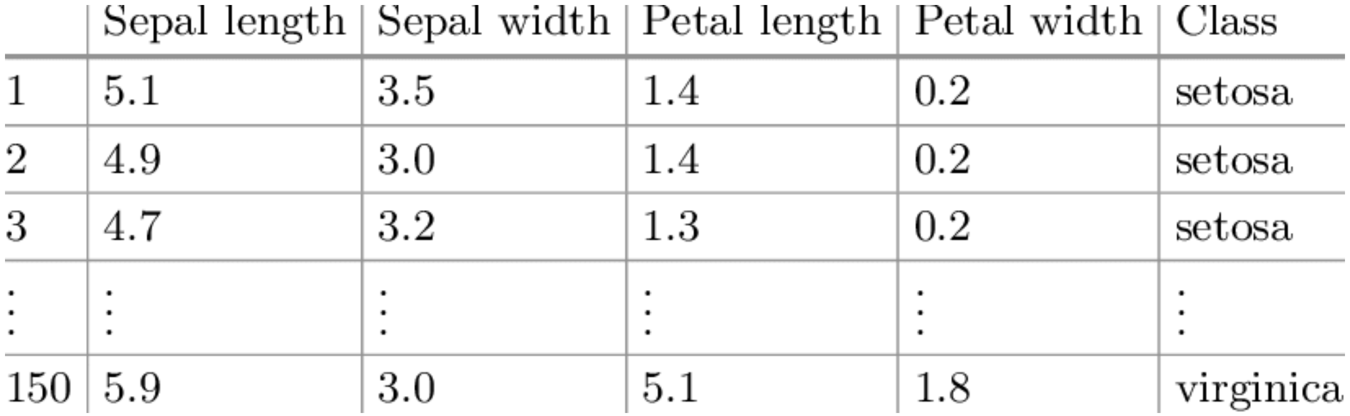
- deals with features of the data set, which form an important input of any machine learning problem

- critical preparatory process in machine learning

- It is responsible for taking raw input data and converting that to well-aligned features which are ready to be used by the machine learning models.

What is feature?

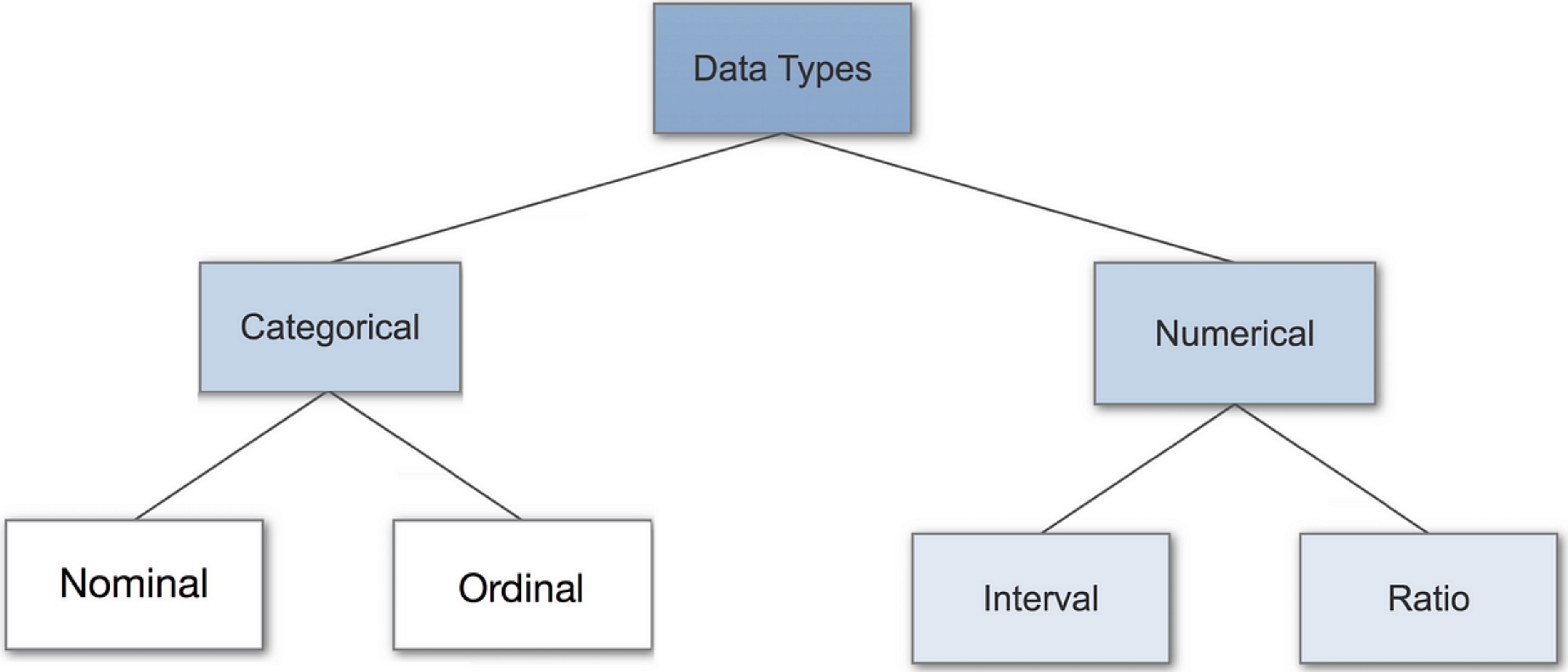
- an attribute of a data set that is used in a machine learning process.
- The features in a data set are also called its dimensions.
- So a data set having ‘n’ features is called an n-dimensional data set.



	Sepal length	Sepal width	Petal length	Petal width	Class
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
⋮	⋮	⋮	⋮	⋮	⋮
150	5.9	3.0	5.1	1.8	virginica

Eg. Iris datatset

Data types in ML



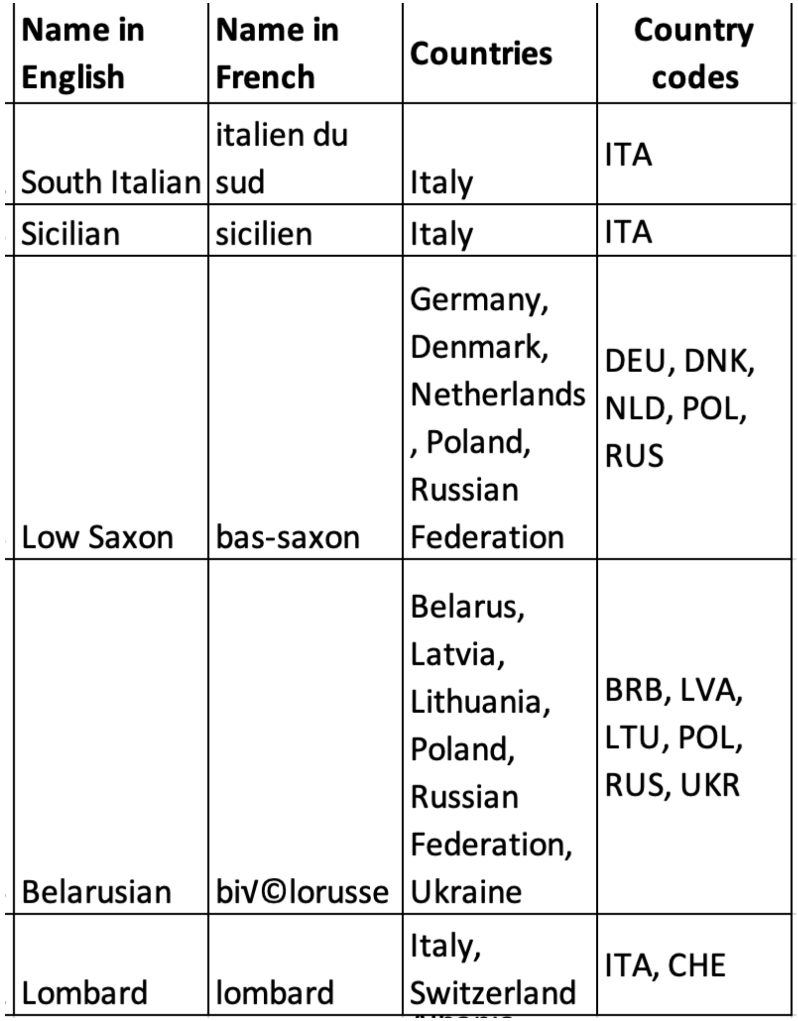
- Function:
  - Features: Inputs to the model, providing the data from which the model learns patterns.
  - Labels: Data Type:
  - Features: Can be numerical, categorical, ordinal, or binary, representing various aspects of the data.
  - Labels: Role in Learning:
  - Features:

Feature vs label

	Feature	Label
Function	Inputs to the model, providing the data from which the model learns patterns.	Outputs the model is trying to predict, serving as the basis for training and evaluating the model.
Data types	Can be numerical, categorical, ordinal, or binary, representing various aspects of the data.	Can be numerical or categorical, representing the outcome or target variable.
Role in Learning	Determine the model's ability to learn; the quality and relevance of features impact model performance.	Guide the learning process; the model uses labels during training to understand how features relate to the outcome.

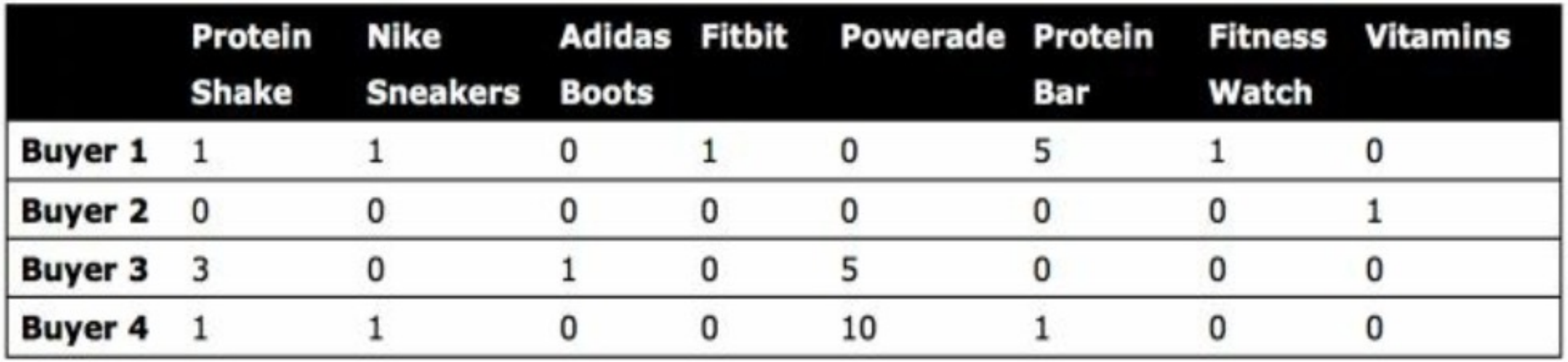
Feature Engineering

- To generate the best results from your data, it is important to first identify the variables most relevant to your hypothesis
- goal is to simplify and speed up data transformations, as well as enhance machine learning model accuracy.
- preserving features that do not correlate strongly with the outcome value can, in fact, manipulate and derail the model’s accuracy.

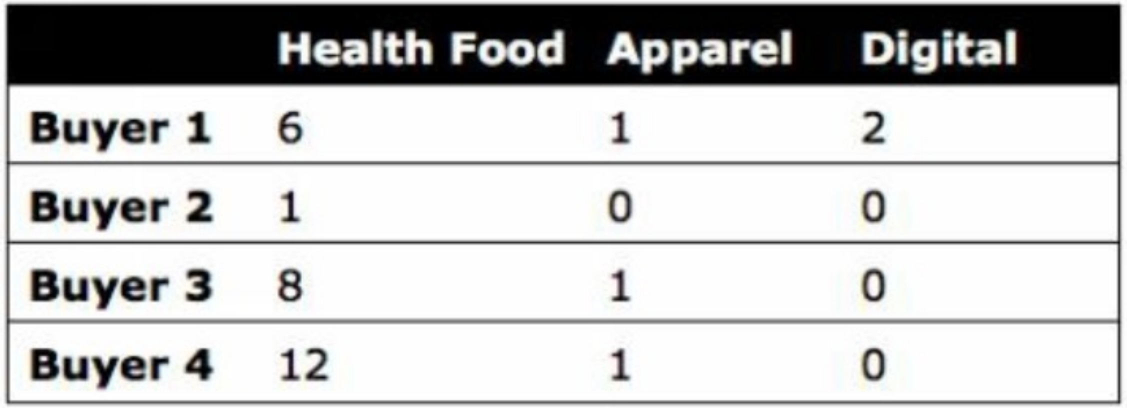


Name in English	Name in French	Countries	Country codes
South Italian	italien du sud	Italy	ITA
Sicilian	sicilien	Italy	ITA
Low Saxon	bas-saxon	Germany, Denmark, Netherlands, Poland, Russian Federation	DEU, DNK, NLD, POL, RUS
Belarusian	biv©lorusse	Belarus, Latvia, Lithuania, Poland, Russian Federation, Ukraine	BRB, LVA, LTU, POL, RUS, UKR
Lombard	lombard	Italy, Switzerland	ITA, CHE

Feature Engineering



	Protein Shake	Nike Sneakers	Adidas Boots	Fitbit	Powerade	Protein Bar	Fitness Watch	Vitamins
Buyer 1	1	1	0	1	0	5	1	0
Buyer 2	0	0	0	0	0	0	0	1
Buyer 3	3	0	1	0	5	0	0	0
Buyer 4	1	1	0	0	10	1	0	0



	Health Food	Apparel	Digital
Buyer 1	6	1	2
Buyer 2	1	0	0
Buyer 3	8	1	0
Buyer 4	12	1	0

Feature Engineering

- the process of translating a data set into features such that these features are able to represent the data set more effectively and result in a better learning performance.
- Two major elements
  1. Feature transformation
     - Feature construction
     - Feature extraction
  2. Feature selection

Data Scrubbing vs feature engineering

- also known as data cleaning,
  - is the process of identifying and correcting (or removing) errors and inconsistencies in the data to improve its quality.
- **Common Tasks:**
  - Handling missing values.
  - Removing duplicates.
  - Correcting data entry errors (e.g., typos, incorrect formats).
  - Dealing with outliers and noisy data.
  - Standardizing data formats (e.g., date formats, text case).
  - Handling inconsistencies in data (e.g., mismatched categories).
- **Goal:** The primary goal of data scrubbing is to ensure that the dataset is clean, consistent, and free of errors so that it doesn't negatively impact the performance of machine learning models.

Data Scrubbing vs **feature engineering**

- the process of selecting, modifying, or creating new features (variables) from raw data to improve the performance of machine learning models.
- **Common Tasks:**
  - Creating new features based on existing data (e.g., polynomial features, interaction terms).
  - Transforming features (e.g., scaling, normalization, log transformations).
  - Encoding categorical variables (e.g., one-hot encoding, label encoding).
  - Extracting features from unstructured data (e.g., text, images).
  - Selecting the most relevant features for the model (feature selection).
- **Goal:** The primary goal of feature engineering is to enhance the predictive power of machine learning models by creating features that better represent the underlying patterns in the data.

Relationship between Data scrubbing and Feature engineering

- **Sequence:**
  - Data scrubbing usually precedes feature engineering.
  - You first need to clean the data to ensure it is reliable before you can effectively engineer features.

- **Dependency:**
  - Clean data is essential for effective feature engineering.
  - If the data is not properly scrubbed, the features engineered from it may be flawed, leading to poor model performance.

Lets make it Clear!!!

- We have learned
  - Splitting data set into two subsets
  - Feature engineering

Which of the above goes first and why ???

Importance of Feature Engineering

- Improve model performance:
  - features are key to the optimal performance of machine learning models.
  - can be thought of as the recipe and the output of the model can be thought of as the meal
- Lessen computational costs:
  - results in reduced computational requirements, like storage, and can improve the user experience by reducing latency
- Improve model interpretability:
  - Speaking to a human's ability to predict a machine learning model's outcomes, well-chosen features can assist with interpretability by helping explain why a model is marking certain predictions

Thank you