LOVELY PROFESSIONAL UNIVERSITY

INT354
Machine Learning

Data Preprocessing

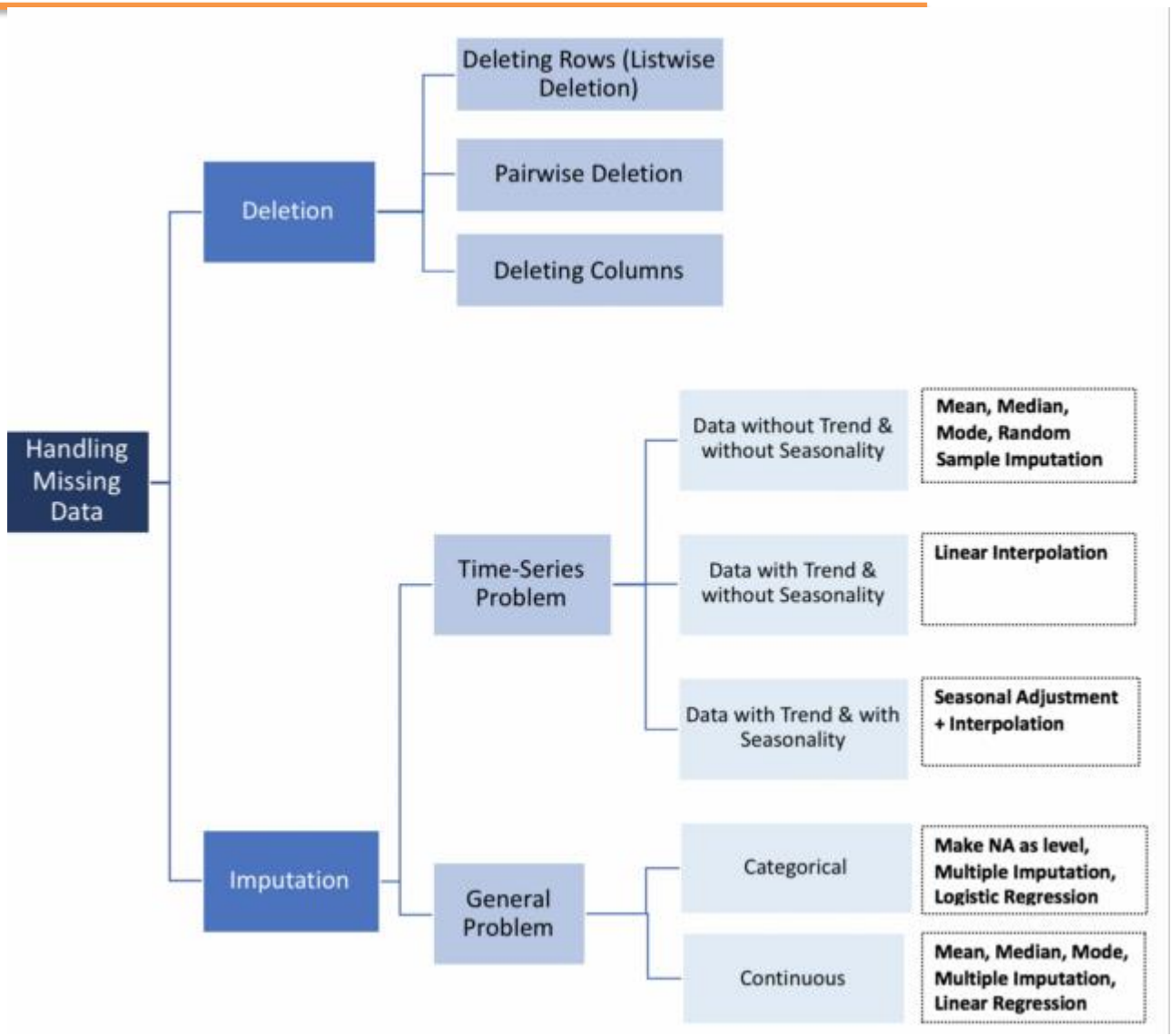
Data Preprocessing

- **Dealing with missing data.**
- **Handling categorical data.**

Ethnicity	Height (CM)	Weight (Kg)	Will Survive till 70
White	186	90	Yes
African	185	98	No
Asian	175	80	No
African	180	88	Yes
Asian	178		No
Asian	172	72	Yes
White	178	75	No
White		89	Yes
African	186	90	Yes

Dealing with missing data

- **An error in the data collection process leads to missing data.**
- **Missing values are represented as *blank spaces* or *NaN* string.**



Eliminating samples or features with missing values

- **Remove the column entirely with missing values.**
- **Remove the rows with missing values.**
- **Only drop rows where all columns are NaN.**
- **Drop rows that have not at least “threshold” non-NaN values.**
- **Only drop rows where NaN appear in specific columns.**

Imputing missing data

- **Removal of entire row/column may lose too much valuable data.**
- **Different interpolation techniques can be used to estimate the missing values like mean/median/mode imputation.**
- **Mean Imputation: replace the missing value by the mean value of the entire feature column.**

Handling Categorical data

- **Mapping ordinal features.**
- **Encoding class labels.**
- **Performing one-hot encoding on nominal features.**

Mapping Ordinal Features

- **Ordinal features are sorted or ordered.**
- **For example:**
  - **Size of T-Shirt: XL>L>M**
- **Convert string values into integer.**
- **For example:**
  - **XL=L+1=M+2**

Encoding Class Labels

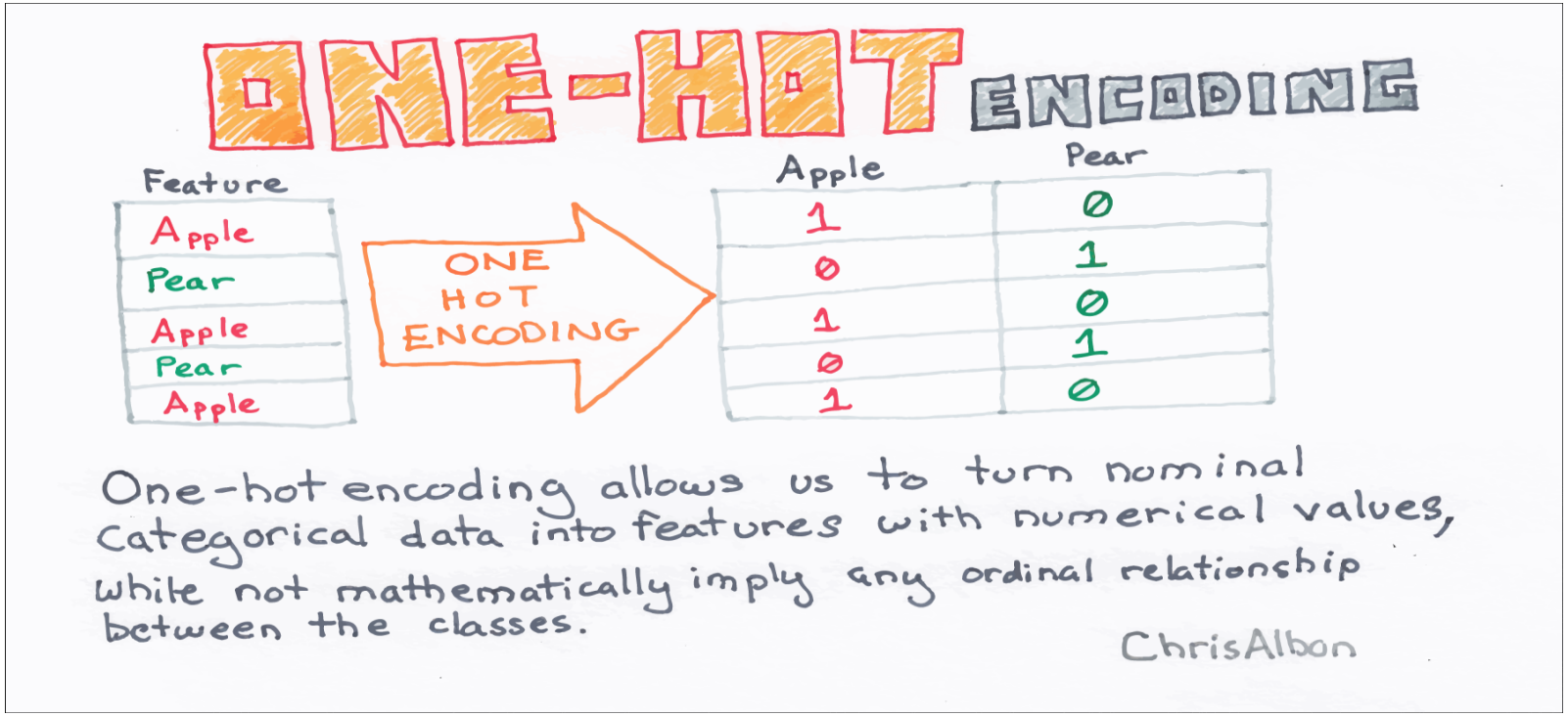
- **Nominal features are not ordered.**
- **For example:**
  - **Color of T-Shirt: green, red, blue.**
- **Assign numeric value to each feature.**
- **For example:**
  - **green: 0, red: 1, blue: 2**

Disadvantages of encoding class labels

- **Assigning numerical values to class labels may lead to wrong interpretation of data.**
- **For example:**
  - **color coding shows that red is greater than green.**

One-hot encoding on nominal features

- **Create a new dummy feature for each unique value.**
- **Assign binary values to indicate a particular feature.**
- **OneHotEncoder returns a sparse matrix.**



Label Encoding vs One Hot Encoding

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Partitioning a dataset

Training Data: it assists in learning and forming a predictive hypothesis for future data.

Test Data: data provided to test a hypothesis created via prior learning is known as test data.

Validation Data: it is a dataset used to retest the hypothesis.


COMING UP