Data Preprocessing with Python

- Data preprocessing-process of preparing the **raw data** and making it suitable for a machine learning model.
- It is the first and crucial step while creating a machine learning model.
- A real-world data generally contains **noises, missing values**, and **inconsistent data** which cannot be directly used for machine learning models.

Preprocessing techniques

1. Encoding categorical features

Majority of the ML algorithms can only work with numerical data therefore categorical variables must be converted to a numerical representation.

Ordinal Encoding

Grades
A
B
C
D
Fail

Grades	Encoded
A	4
B	3
C	2
D	1
Fail	0

Red
Blue
Green

Red	Blue	Green
1	0	0
0	1	0
0	0	1

2. Imputing missing values

Most machine learning algorithms cannot interpret null values.

glucose	BP	isulin	BMI	age	class variable
148	72	0	33.6	50	tested_positive
85		0	26.6	31	tested_negative
183	64	0	23.3	32	tested_positive
89	66	94	28.1	21	tested_negative
137	40	168	43.1	33	tested_positive
116		0	25.6	30	tested_negative
78	50	88	31	26	
115	0	0	35.3	29	tested_negative
197		543	30.5	53	tested_positive
125	96	0	0	54	tested_positive
110	92	0	37.6	30	tested_negative

Preprocessing techniques cont.

3. Feature scaling

ML algorithms only understand numerical relationships. Features with varying scales may therefore be incorrectly interpreted.

Values	Normalized	Standardized
47	0.9302	1.1560
7	0.0000	-1.9267
21	0.3256	-0.8478
28	0.4884	-0.3083
41	0.7907	0.6936
49	0.9767	1.3102
50	1.0000	1.3872
25	0.4186	-0.5395
25	0.4186	-0.5395
35	0.6512	0.2312
24	0.3953	-0.6165

4. Binning

Continuous variables with many infrequently occurring values can contain a lot of noise which might lead to overfitting.

Binning aggregates these values into groups of similar values resulting in a new categorical feature.

Data preprocessing stages

- Acquire the dataset
- Import all the crucial libraries
- Import the dataset
- Identifying and handling the missing values
- Encoding the categorical data (if any)
- Splitting the dataset
- Feature scaling

Importing the libraries

#Importing the libraries: a Library is a tool to make a specific job/function

import numpy as np *# numpy facilitates any type of mathematical operation in the code*

import matplotlib.pyplot as plt *# pyplot is a sublibrary, plots nice charts*

import pandas as pd *# facilitate data imports and data management*

import os *# for displaying the current working directory*

The current working directory

 # check for the current working directory

cwd=os.getcwd()

#Setting the working directory insure the datafile and the python file (.py) are in the same folder then #Execute the file or press F5

#os.chdir('D:/MACHINE LEARNING')

Importing datasets

```
# Importing datasets
datasets=pd.read_csv('Data.csv') #reads a csv file in python
#datas=pd.read_stata('TNA2.dta')
```