240324_8기_데이터전처리 (판다스)

Missing Value

Missing Data Mechanisms

- MCAR – *p*(missing) is *unrelated* to all variables, observed and unobserved

  $$p(\text{missing}|\text{complete data}) = p(\text{missing})$$

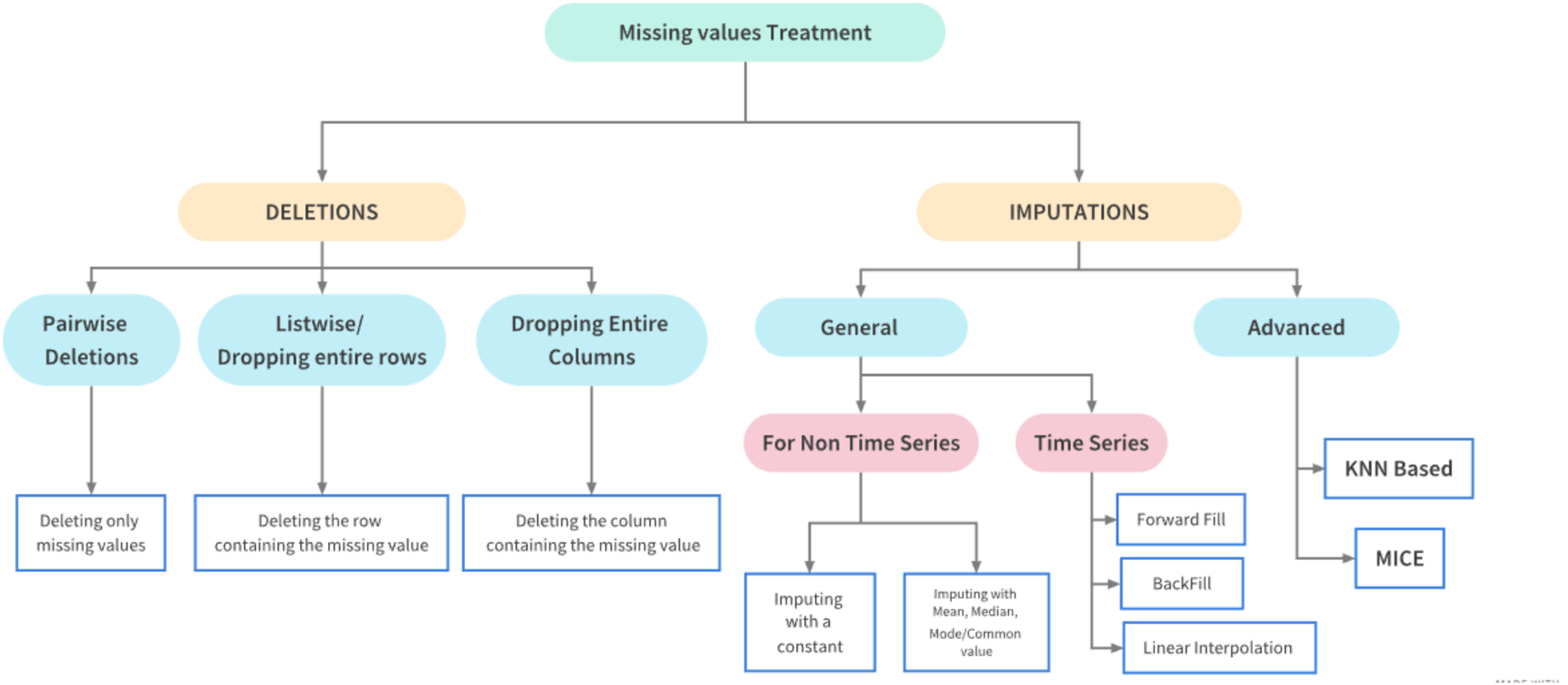
- MAR – *p*(missing) is related to observed variables [observed data] only

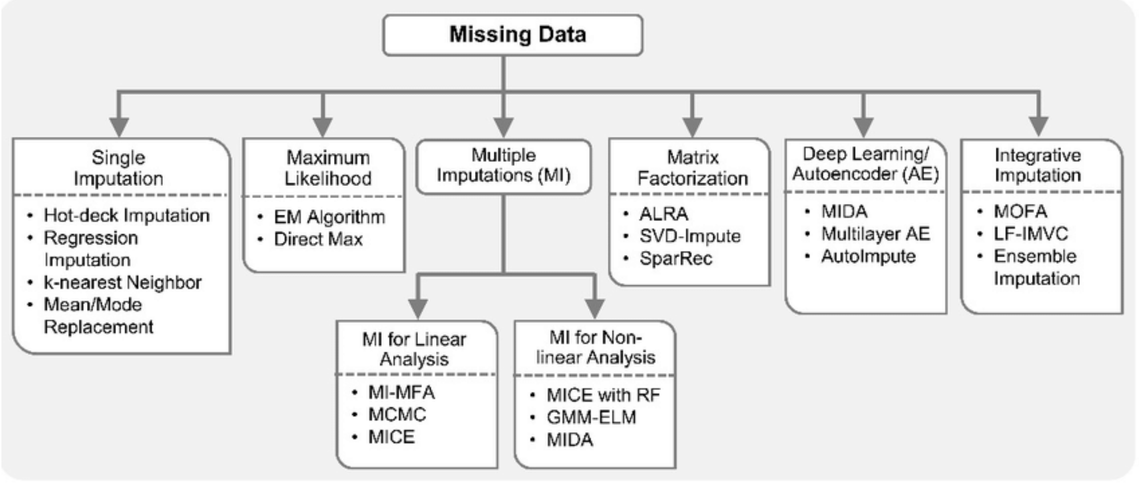
  $$p(\text{missing}|\text{complete data}) = p(\text{missing}|\text{observed data})$$

- MNAR – *p*(missing) is related to the unobserved/ missing variables [missing data]

  $$p(\text{missing}|\text{complete data}) \neq p(\text{missing}|\text{observed data})$$

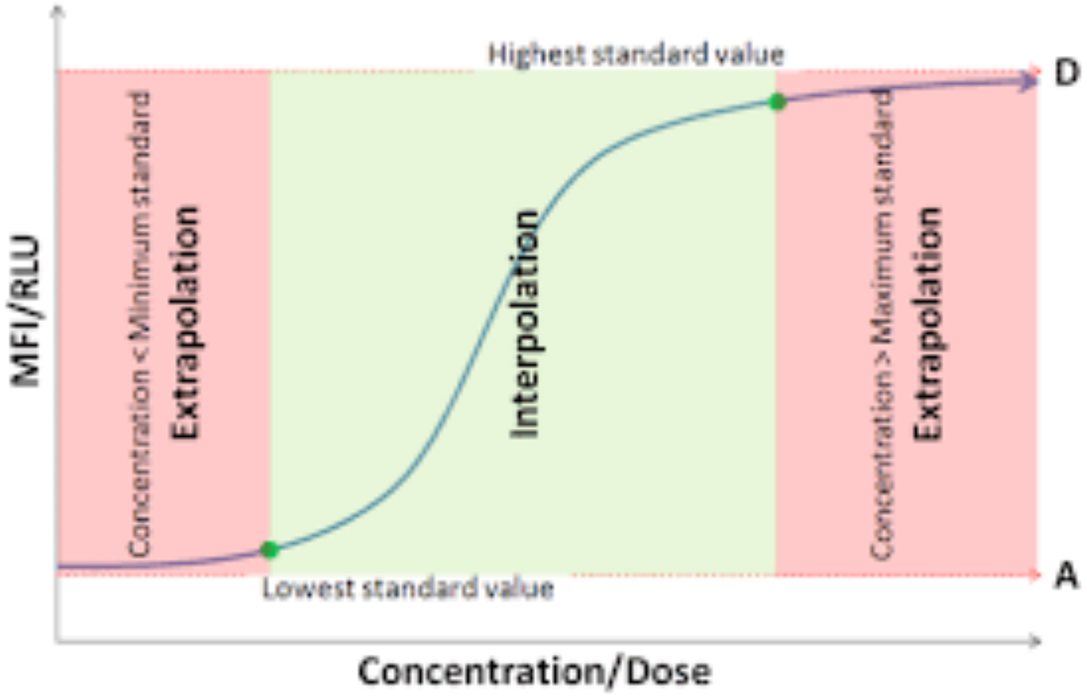
(see Schafer & Graham, 2002)

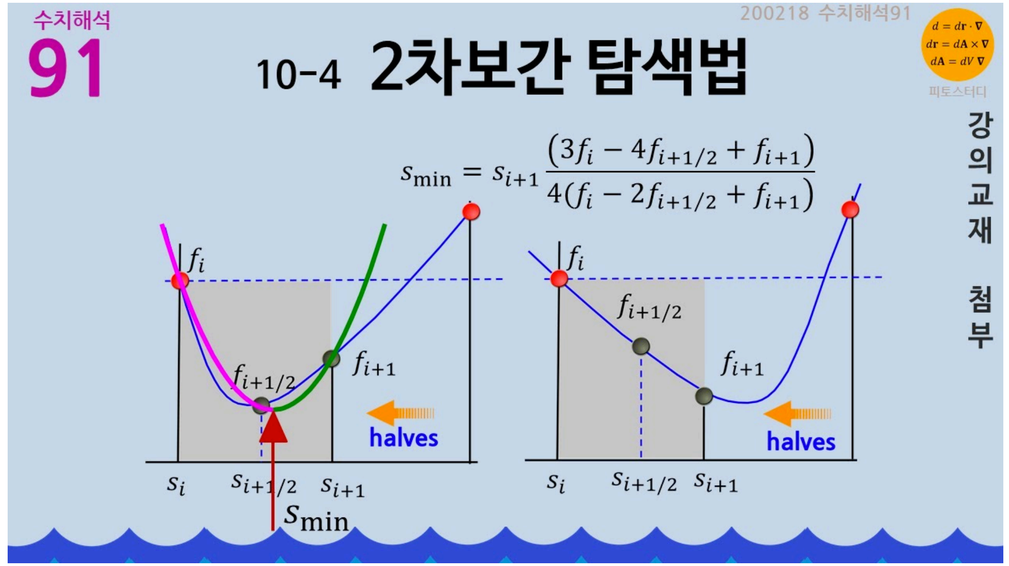

Missing values Treatment
DELETIONS
IMPUTATIONS
Pairwise Deletions
Listwise/ Dropping entire rows
Dropping Entire Columns
General
Advanced
Deleting only missing values
Deleting the row containing the missing value
Deleting the column containing the missing value
For Non Time Series
Time Series
KNN Based
MICE
Forward Fill
BackFill
Linear Interpolation
Imputing with a constant
Imputing with Mean, Median, Mode/Common value


Missing Data
Single Imputation
• Hot-deck Imputation
• Regression Imputation
• k-nearest Neighbor
• Mean/Mode Replacement
Maximum Likelihood
• EM Algorithm
• Direct Max
Multiple Imputations (MI)
Matrix Factorization
• ALRA
• SVD-Impute
• SparRec
Deep Learning/ Autoencoder (AE)
• MIDA
• Multilayer AE
• AutoImpute
Integrative Imputation
• MOFA
• LF-IMVC
• Ensemble Imputation
MI for Linear Analysis
• MI-MFA
• MCMC
• MICE
MI for Non-linear Analysis
• MICE with RF
• GMM-ELM
• MIDA

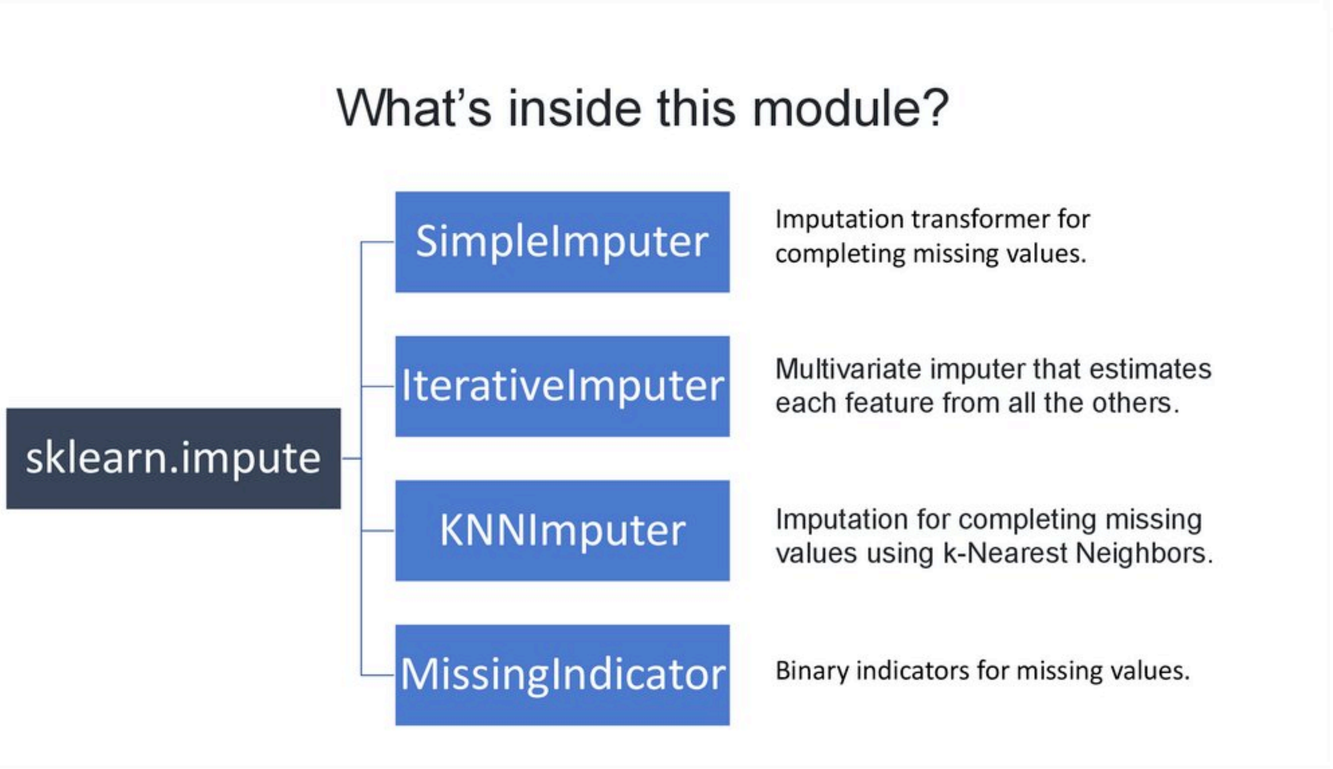
Interpolation

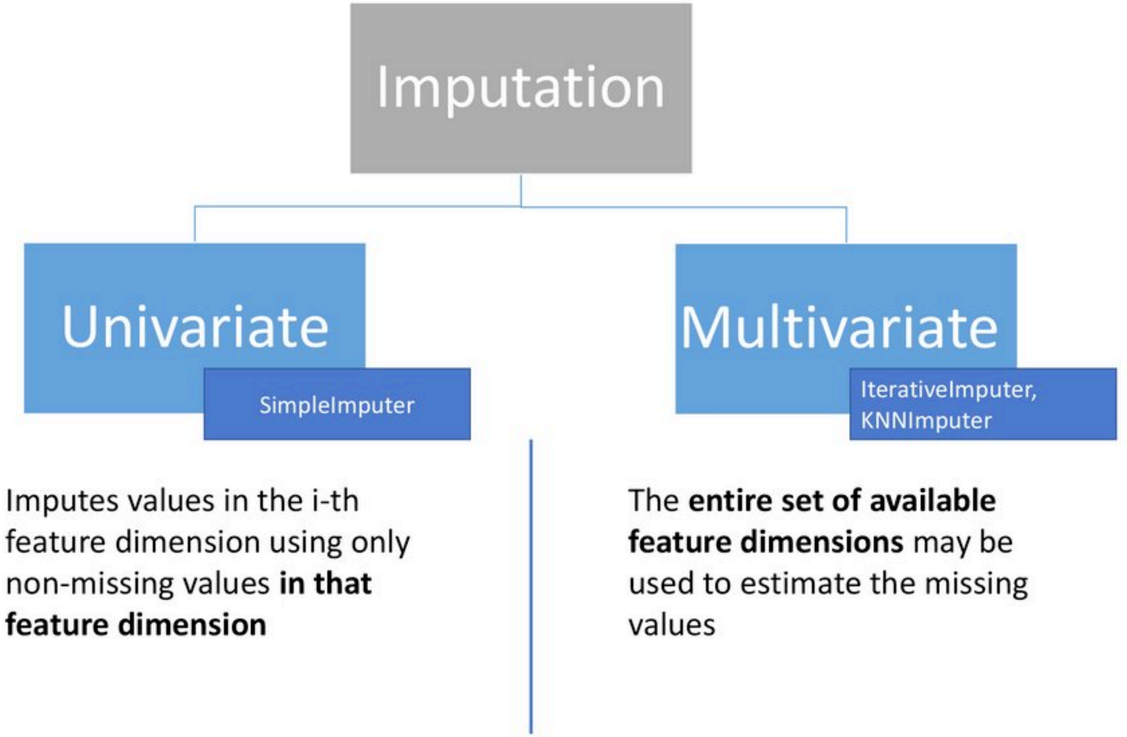
- 보간법 (Interpolation)
- 내삽이라 하며 특정한 두 점 안쪽에 놓여있는 가능한 값을 구하려는 방법
- 보외법 (Extrapolation)
- 외삽 관찰 범위를 넘어서 다른 변수와의 관계에 기초하여 변수의 값을 추정하는 과정





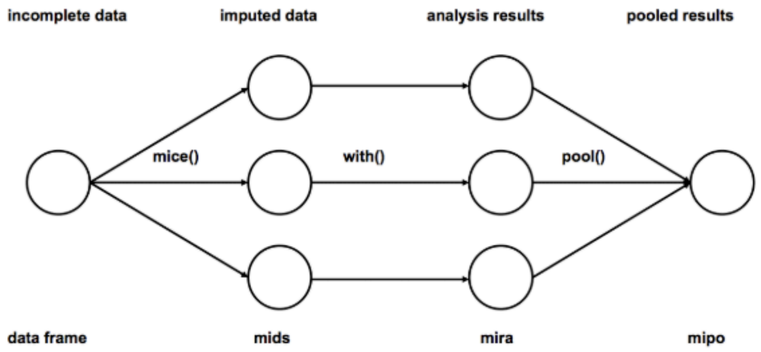
Sklearn impute





Iterative Imputer

1. **Imputation**: Impute the missing entries of the incomplete data sets m times (m=3 in the figure). Note that imputed values are drawn from a distribution. Simulating random draws doesn't include uncertainty in model parameters. Better approach is to use Markov Chain Monte Carl (MCMC) simulation. This step results in m complete data sets.
2. **Analysis**: Analyze each of the m completed data sets.
3. **Pooling**: Integrate the m analysis results into a final result



Multiple Imputation 3단계

- Imputation Phase: 가능한 대체 값의 분포에서 추출된 서로 다른 값으로 복수의 데이터 셋을 생성
- Analysis Phase: 각 데이터 셋에 대하여 모수의 추정치와 표본오차 계산
- Pooling Phase: 모든 데이터 셋의 추정치와 표본오차를 통합하여 하나의 대치값 생성