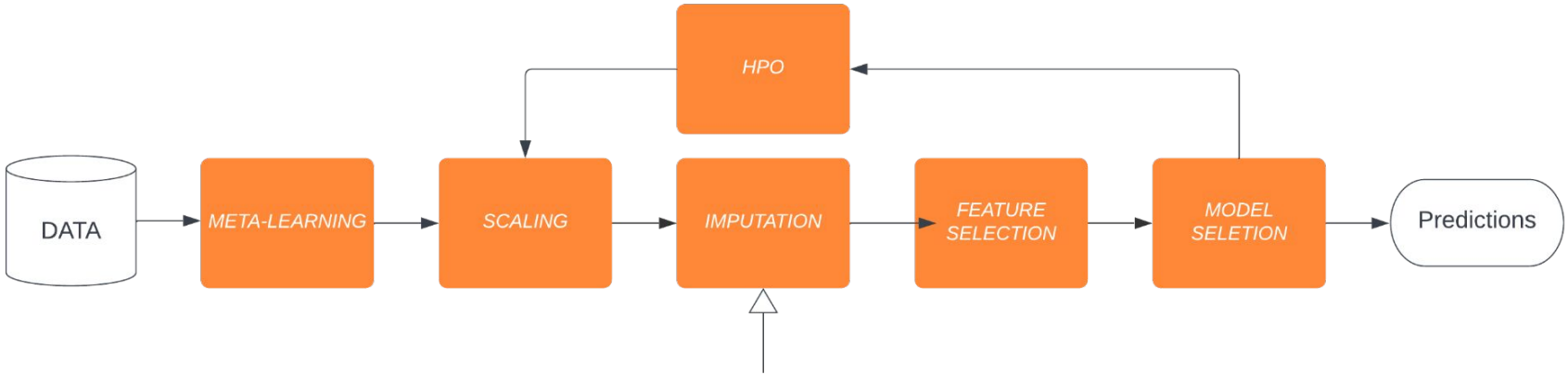
Missing data

Udemy course

Why is it a problem?

- Many machine learning algorithms require complete data sets.
- Different types of missing data may require different methods.
- Mostly studied by statisticians and not by ML scientists.
- Literature reports contradicting results.

Imputation in the AutoML pipeline



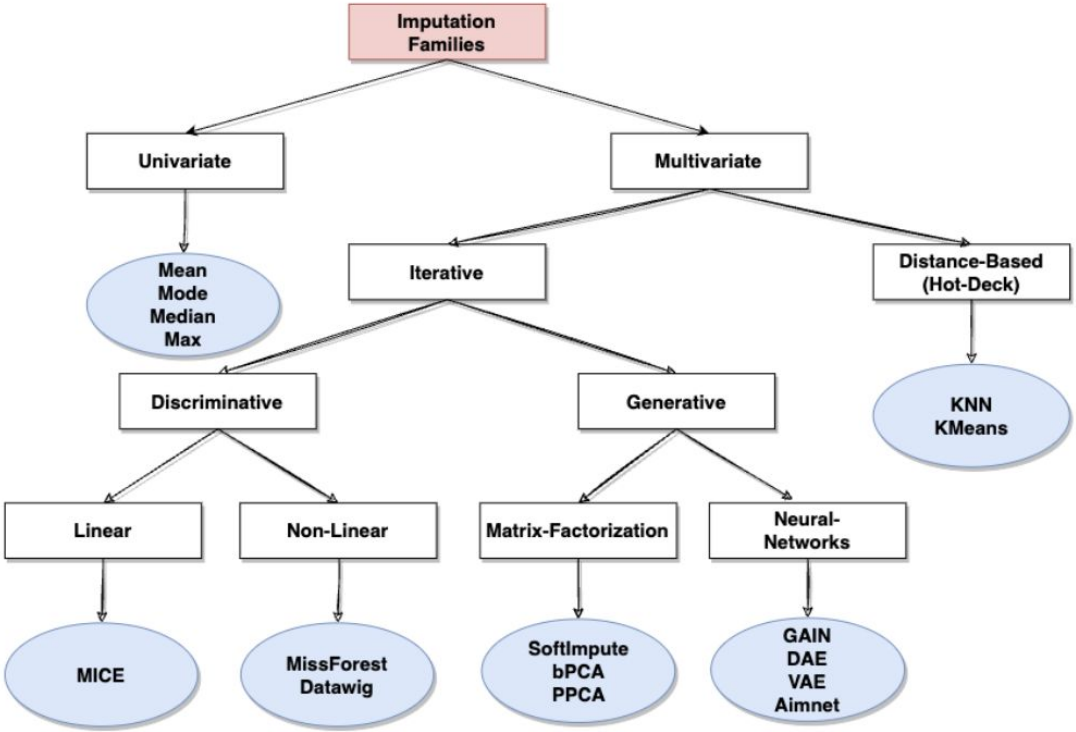
Imputation in AutoML tools

Tool	Simple-Imp.	SoftImpute	MissForest	New cat.	Other
JadBio	✔				
TPOT	✔				
AutoPrognosis	✔	✔	✔		✔
AutoGluon	✔			✔	
AutoSklearn	✔				✔

Missingness mechanisms

- Missing Completely at Random (MCAR): Missing values don't depend on the data.
  - E.g : A dataset got corrupted when compressed and some values were lost.
- Missing at Random (MAR) : Missing values are related to the other observed variables.
  - E.g: People with higher income may to not answer how much taxes they pay.
- Missing Not at Random (MNAR) : Missing values are related to the values of the variable itself.
  - E.g : People with higher income may refuse to answer how much money they make.

Taxonomy of Imputation methods



Ref: Do we really need imputation in AutoML predictive modeling? George Paterakis et al. 2023

Mean / Mode Imputation

- Mean Imputation :
  - Used for numeric features that have missing values.
  - Replace missing values with the mean value of the feature.
- Mode Imputation :
  - Used for categorical features that have missing values
  - Replace missing values with most frequent feature value.

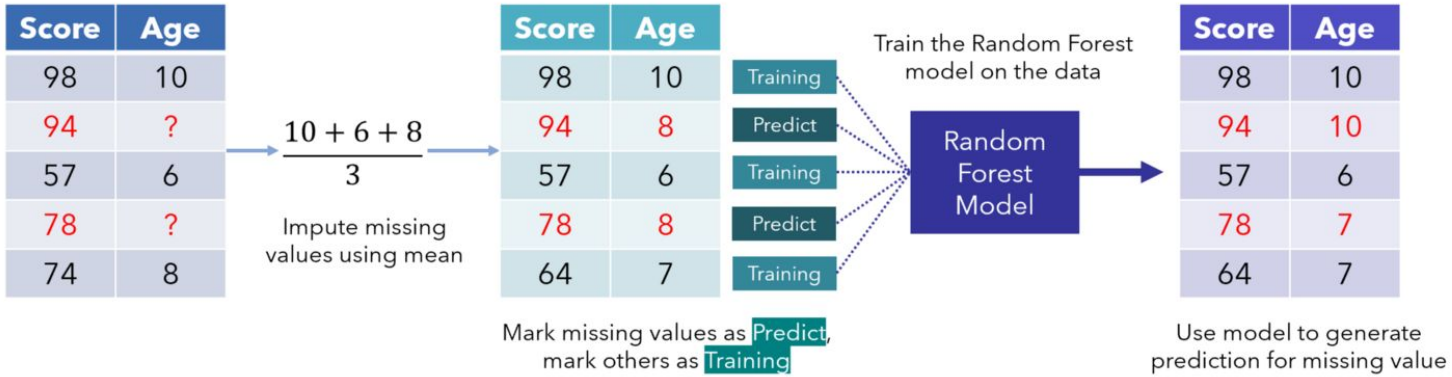
X1	X2	Y
60	'NO'	10
80	'YES'	15
-	'YES'	10
100	-	15

X1	X2	Y
60	'NO'	10
80	'YES'	15
80	'YES'	10
100	'YES'	15

MissForest

Iterative imputation based on Random Forest.

- First the missing values are imputed by mean/mode.
- Then for each feature with missing values, we train a RF on the other features and predict the missing values using the trained RF.
- Continue iteratively using the last imputed dataset till convergence.



SoftImpute

SoftImpute is an iterative method utilizing a rank-restricted SVD.

- Missing values are zero-imputed
- In each iteration a rank restricted SVD is applied, then the data are projected using the calculated matrices to the original dimensions.
- Continue iteratively using the last imputed dataset till convergence.

Probabilistic PCA

Probabilistic PCA is an iterative method utilizing a generalized PCA.

- Missing values are zero imputed
- In each iteration, PPCA projects data in a lower dimension and reconstructs the data using the learned parameters.
- Continue iteratively using the last imputed dataset.

PPCA assumptions:

Linear generative model: $y_d = \sum_{k=1} \Lambda_{dk}\, x_k + \epsilon_d$

- x_k are independent $\mathcal{N}(0, 1)$ Gaussian factors
- ϵ_d are independent $\mathcal{N}(0, \sigma^2)$ Gaussian noise
- $K < D$

PPCA is factor analysis with isotropic noise: $\Psi = \sigma^2 I$

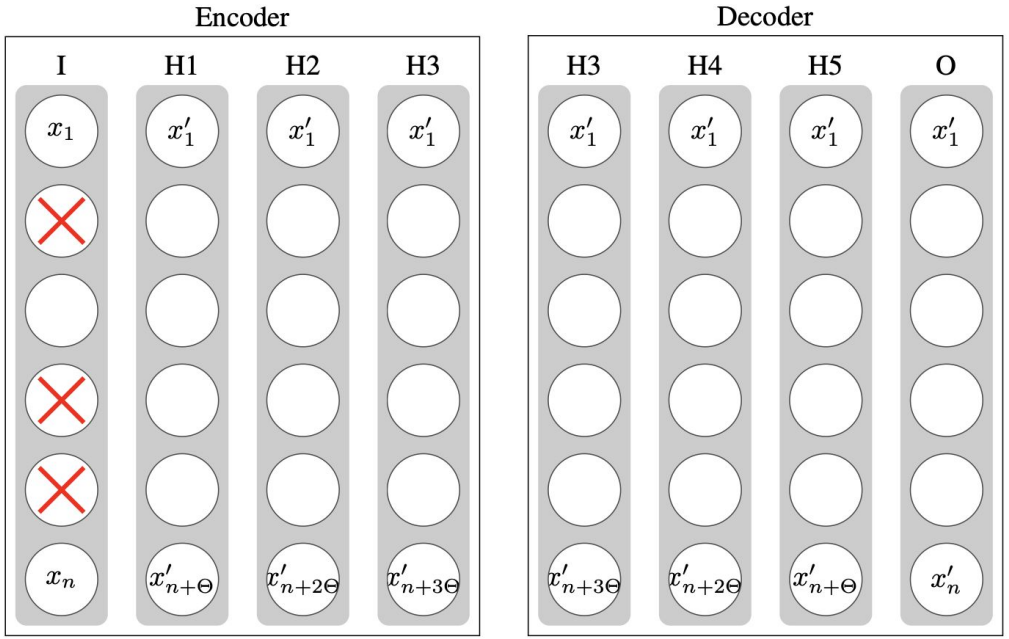
Denoise Autoencoder

Denoising autoencoders is a neural network that contains an encoder and a decoder.

- First it does an initial imputation using mean/mode.
- Then it corrupts the input data by setting a % input of data equal to 0. (Dropout-Layer)
- It sequentially it maps data to a higher dimensional space. (Encoder)
- Then it sequentially maps the data to a lower dimensional space till it reaches the original dimension. (Decoder)

Overview of DAE.

We start with an initial n dimensional input, then at each successive hidden layer, we add Θ nodes. Then the decoder maps data to -Θ dimension sequentially.

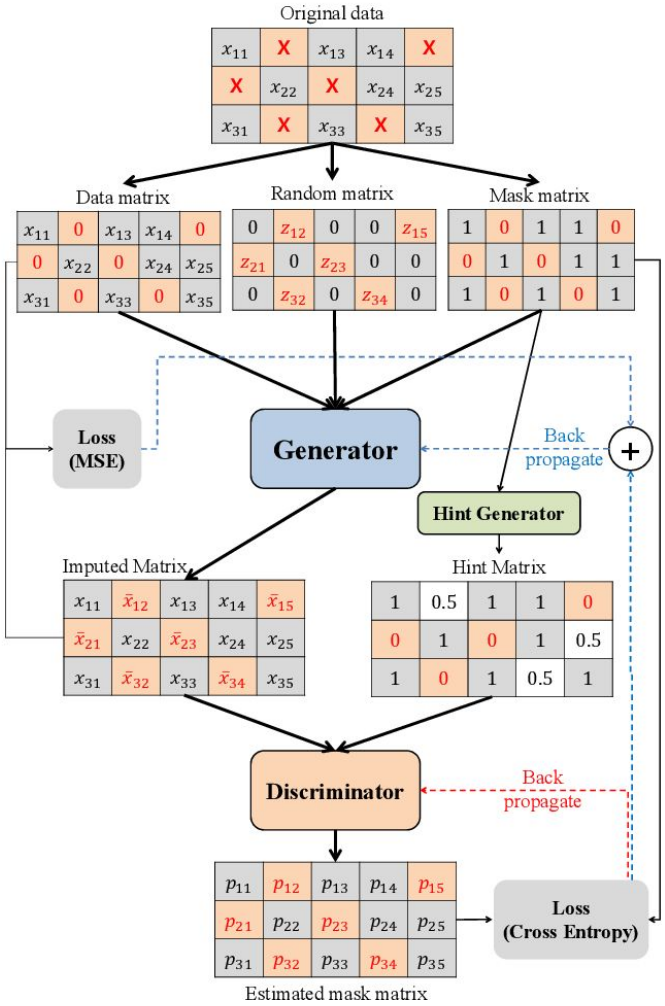


GAIN

GAIN is a generalization of GAN for the purpose of missing data imputation.

Generator : Tries to accurately impute missing data.

Discriminator: Predict which data are observed and which are imputed given a hint matrix.



Missing Indicator

- Missing Indicator Method :
  - Impute missing values with a method.
  - Create a new variable for every feature with missing values.
  - The new variable takes value 1 if value is missing , 0 if not missing.

X1	Y
60	10
80	15
-	10
100	15

X1	X1R	Y
60	0	10
80	0	15
80	1	10
100	0	15

Hyper-parameters

Method	Hyper Parameter	Values
MM	-	-
MissForest	max_depth	10, 20
SoftImpute	Variance explained	50%, 70%, 90%
PPCA		
DAE	Dropout	0.25, 0.4, 0.5
	theta	5, 7, 10
GAIN	hint_rate	0.5, 0.9
	alpha	0.1, 1, 10