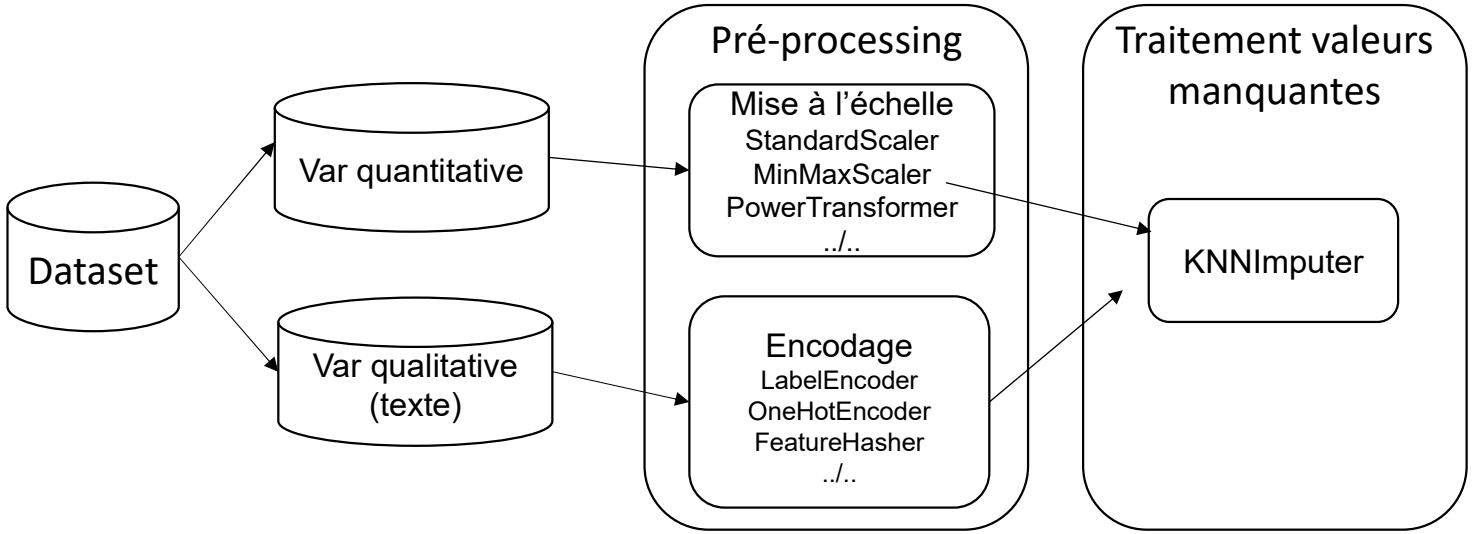
1 Imputation

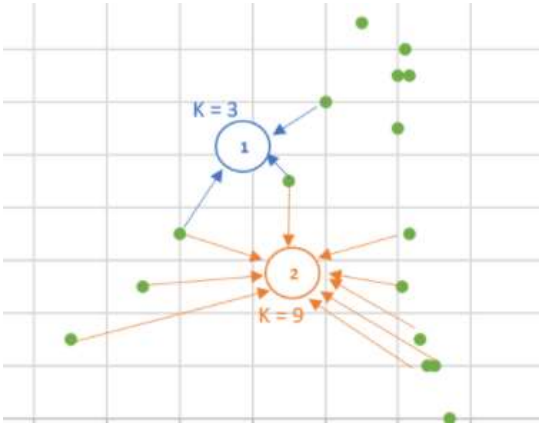
1. Imputation - KNNImputer

Basé sur l'algorithme K-Nearest Neighbors

Pour chaque valeur manquante d'un point de données :

1. KNN Imputer **cartographie** l'ensemble de données à l'exclusion des éléments ayant des valeurs manquantes dans l'espace de coordonnées à n dimensions
2. **Calcule** la **distance** euclidienne (par défaut) des **points** les **plus proches** de ce point de données.
3. **Imputation** des valeurs manquantes par la **moyenne** des éléments pertinents pour ces points les plus proches





1. Imputation - KNNImputer

- Encodage pour var. qualitatives.
- Mise à l'échelle pour les variables quantitatives.
- Inversion de la mise à l'échelle pour avoir les vraies valeurs.
- Basé sur hyper-paramètre k ➔ tri-it-all
- Hypothèse de relations entre les entités
- Mauvaises prédictions si prédicteurs faibles ou fortes relations entre les entités
- Sujet à la malédiction (fléau) de la dimensionnalité

https://ichi.pro/fr/missforest-le-meilleur-algorithme-d-imputation-des-donnees-manquantes-5291693485779

- Imputer des valeurs autres que constantes (numériques ou modalités, moyenne, mode ou médiane.

NaNimputer?
Verstack

MissForest ?
Appliquée sur données mixtes,
Encodage, pas mise à l'échelle
Pas d'hypothèse de relation entre entités
Robuste aux bruits et à la mutlicolinéarité
Non paramétrique : pas de réglages
Grande dimension
Mais plus long petits datasets