Résumé sur les étapes principales de prétraitement

Qu'est-ce que le prétraitement des données ?

La plupart des données disponibles par défaut sont trop brutes. Il est important de prétraiter les données avant de pouvoir les utiliser pour identifier des modèles ou peuvent être utilisés pour former des modèles statistiques qui peuvent être utilisé pour faire des prédictions. Le prétraitement des données est le processus de nettoyage et d'ingénierie donnée de manière à ce qu'elles puissent être utilisées comme données d'entrée pour plusieurs tâches de science des données telles que, l'apprentissage en profondeur dans notre cas.

Certaines des tâches de prétraitement de données les plus courantes incluent la détection des valeurs aberrantes, le traitement des valeurs manquantes, l'encodage des variables qualitatives, la normalisation, la discrétisation des données, etc.

1. Traitement des valeurs manquantes (nettoyage des données)

1.1. Méthodes avec suppression de données

Dans certains cas, l'analyse est possible sans imputer les données manquantes. En général, on se reporte à deux méthodes classiques :

- **L'analyse des cas concrets** : qui consiste à ne considérer que les individus pour lesquels toutes les données sont disponibles, i.e. en supprimant les lignes comportant des valeurs manquantes. C'est ce qui est fait automatiquement avec R (na.action=na.omit). Cette méthode, risque de supprimer trop de données et d'augmenter de beaucoup la perte de précision.

- **L'analyse des cas disponibles** : Afin d'éviter de supprimer trop de données, il est possible de faire de la suppression par paires (pairwise deletion) ou analyse des cas disponibles (available-case analysis). Différents aspects du problème sont alors étudiés avec différents sous échantillons. Cependant, les différentes analyses ne seront pas nécessairement compatibles entre elles. L'analyse des cas disponibles correspond aussi au cas où une variable est supprimée du jeu de données à cause de sa trop grande quantité de valeurs manquantes.

1.2. Méthodes d'imputation :

❖ **Imputation moyenne ou médiane :** technique où les valeurs manquantes dans une colonne sont remplacées par la moyenne ou la médiane de toutes les valeurs restantes dans cette colonne particulière.

- Ces imputations sont faciles à mettre en œuvre et constituent une stratégie utile pour obtenir rapidement un grand ensemble de données. De plus, possibilité de mises en œuvre lors de la phase de production.
- Son plus grand inconvénient est qu'elle affecte la distribution des données par défaut ainsi que la variance et la covariance des données.

❖ **Imputation de fin de distribution** Pour les données aléatoirement manquantes, les techniques les plus couramment utilisées sont l'imputation de fin de distribution/fin de queue. À la fin de l'imputation de queue, une valeur est choisie à partir de la fin des données. Cette valeur signifie que les données réelles de l'enregistrement étaient manquantes. Par conséquent, les données qui ne manquent pas au hasard peuvent être prises en compte lors de la formation de modèles statistiques sur les données.

- L'un des principaux avantages de l'imputation de fin de distribution est qu'elle peut être appliquée à l'ensemble de données où les valeurs ne manquent pas au hasard. Les autres avantages incluent sa simplicité de compréhension, sa capacité à créer de grands ensembles de données en peu de temps et son applicabilité dans l'environnement de production.

❖ **Imputation de valeur arbitraire** En fin d'imputation de distribution, les valeurs qui remplacent les valeurs manquantes sont calculées à partir des données, tandis que dans l'imputation de valeurs arbitraires, les valeurs utilisées pour remplacer les valeurs manquantes sont sélectionnées arbitrairement. Les valeurs arbitraires sont sélectionnées de manière à ne pas appartenir à l'ensemble de données ; elles signifient plutôt les valeurs manquantes.
- Il est important de mentionner que l'imputation de valeurs arbitraires peut également être utilisée pour les données catégorielles.

Les deux restants sont pour les données catégorielles

❖ **Imputation par catégorie fréquente** consiste à remplacer les valeurs manquantes par les valeurs les plus fréquentes, c'est-à-dire le mode de la colonne. C'est pour cette raison que l'imputation fréquente par catégorie est également connue.
- L'imputation par catégorie fréquente est plus facile à mettre en œuvre sur grands ensembles de données. La distribution fréquente des catégories ne fait pas toute

hypothèse sur les données et peut être utilisé dans une production environnement.

- **Imputation de catégorie manquante** L'imputation de la catégorie manquante est similaire à la valeur arbitraire. Dans le cas d'une valeur manquante l'imputation ajoute une catégorie arbitraire.

2. Encodage

Pour les données catégorielles (qualitatives)

- **One Hot Encoding :** chaque valeur unique dans une colonne catégorielle, est représentée en forme binaire dans une colonne qui lui est propre tel que : L'entier 1 est ajouté à la colonne qui correspond à l'étiquette d'origine, et toutes les colonnes restantes sont remplies de zéros.

- **Frequency Encoding :** Dans le codage fréquentiel, chaque étiquette unique dans une catégorie est remplacée par son nombre total ou sa fréquence. Par exemple, dans le tableau suivant, USA apparaît trois fois, tandis que UK et La France compte respectivement deux et un.

Country	encodage
UK	2
USA	3
UK	2
USA	3
FRANCE	1
USA	3

- **Ordinal Encoding :** Dans le codage ordinal, on associe chaque catégorie d'une valeur décimale unique. Les étiquettes sont classées sur la base de leur relation avec la cible.

3. Discrétisation

C'est le processus de conversion de valeurs numériques continues comme prix, l'âge et le poids dans des intervalles discrets.

- **Equal Width Discretization** En discrétisation à largeur fixe, la largeur ou la taille de tous les intervalles reste la même. Un intervalle est également appelé bin.
- **Equal Frequency Discretization** Dans la discrétisation à fréquence égale, la largeur de la case est ajustée automatiquement de telle sorte que chaque case

contienne exactement le même nombre d'enregistrements ou ait la même fréquence. D'où le nom de discrétisation à fréquence égale.

Il y a d'autre techniques de discrétisation comme : K-Means Discretization, Decision Tree Discretization, Custom Discretization

4. Traitement des valeurs aberrantes

4.1. Outlier Trimming

Le découpage des valeurs aberrantes, comme son nom l'indique, consiste simplement à supprimer les valeurs aberrantes au-delà d'une certaine valeur seuil. L'un des principaux avantages du découpage des valeurs aberrantes est qu'il est extrêmement rapide et ne déforme pas les données. Un inconvénient du découpage des valeurs aberrantes est qu'il peut réduire la taille des données.

Il existe une autre façon de gérer les valeurs aberrantes

4.2. Outlier Capping

Dans le plafonnement, les valeurs aberrantes sont plafonnées à certaines valeurs minimales et maximales. Les lignes contenant les valeurs aberrantes ne sont pas supprimées du jeu de données.

Outlier capping se fait en utilisant :

a. **IQR** : Nous utiliserons la technique Inter Quartile Range pour trouver la limite inférieure et supérieure des valeurs aberrantes dans les colonnes.
b. **Moyenne and Std** : Au lieu d'utiliser la méthode IQR, les seuils supérieurs et inférieurs pour les valeurs aberrantes peuvent être calculés via la méthode de la moyenne et de l'écart type. Pour trouver le seuil supérieur, la moyenne des données est ajoutée à trois fois la valeur de l'écart type. De même, pour trouver le seuil inférieur, vous devez multiplier l'écart type par 3, puis retirer le résultat de la moyenne
c. **Quantiles** : On peut également utiliser les informations quantiles pour définir les limites inférieure et supérieure afin de trouver des valeurs aberrantes
d. **Valeurs Personnalisés :** Enfin, vous pouvez également définir les valeurs personnalisées pour les limites inférieure et supérieure afin de trouver les valeurs aberrantes.

5. La normalisation

Mettre les données en même échelle pour faciliter l'apprentissage (la convergence de fonction coût)

❖ **Standardisation**

Consiste à centrer la variable sur zéro et normaliser la variance des données à 1. Pour normaliser l'ensemble de données, il vous suffit de soustraire chaque point

de données de la moyenne du point de données et diviser le résultat par l'écart type des données. $Xscaled = \frac{X-\mu X}{\sigma X}$

- **Normalisation Min/Max**
  Dans la mise à l'échelle min/max, vous soustrayez chaque valeur par la valeur minimum, puis divisez le résultat par la différence de minimum et la valeur maximale dans l'ensemble de données. $Xscaled = \frac{X-X_{min}}{X_{max}-X_{min}}$
- **Mean Normalization** La normalisation moyenne est très similaire à la mise à l'échelle min/max, sauf dans la normalisation moyenne, la moyenne de l'ensemble de données est soustraite de chaque valeur, et le résultat est divisé par la plage, c'est-à-dire différence entre les valeurs minimales et maximales. $Xscaled = \frac{X-moyenne}{X_{max}-X_{min}}$
- **Median and Quantile Scaling**
  Dans l'échelle médiane et quantile, la moyenne de l'ensemble de données est soustraite de tous les points de données, et le résultat est divisé par la différence entre le premier quartile et le 3e quartile. $Xscaled = \frac{X-mediane}{IQR}$.