Projet de conception et de programmation

Mickael LE DENMAT

Université Versailles Saint-Quentin en Yvelines
Investigating feature selection techniques to improve data mining tasks

5 février 2023

Table des matières

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Data Mining

- Le monde d'aujourd'hui : Beaucoup de données !

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Data Mining

- Le monde d'aujourd'hui : Beaucoup de données !
- Problème de prise de décision.

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Data Mining

- Le monde d'aujourd'hui : Beaucoup de données !
- Problème de prise de décision.

Solution

Data Mining : la pratique consistant à rechercher automatiquement de grandes quantités de données afin de découvrir des tendances et des modèles qui vont au delà de la simple analyse. [Oracle,]

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Feature selection

- Complexité du monde réel.

Feature selection

- Complexité du monde réel.
- Accélérer la prise de décision.

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Feature selection

- Complexité du monde réel.
- Accélérer la prise de décision.

Solution

Feature selection.

Rough Set Theory

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Indiscernibility Relation

Exemple

Patient	Headache	Muscle-Pain	Temperature	Flu
o_1	Yes	Yes	Very High	Yes
o_2	Yes	No	High	Yes
o_3	Yes	No	High	No
o_4	No	Yes	Normal	No
o_5	No	Yes	High	Yes
o_6	No	Yes	Very High	Yes

Table – Dessin pour le calcul du IND_C

$$IND_C = \{\{o_1\}, \{o_2, o_3\}, \{o_4\}, \{o_5\}, \{o_6\}\}.$$

B-Lower & B-Upper Approximation

Rappel

$X_1 = \{o_j | Flu(o_j) = \{Yes\}\} = \{o_1, o_2, o_5, o_6\}$
$X_2 = \{o_j | Flu(o_j) = \{No\}\} = \{o_3, o_4\}$
$IND_C = \{\{o_1\}, \{o_2, o_3\}, \{o_4\}, \{o_5\}, \{o_6\}\}.$

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

B-Lower & B-Upper Approximation

Rappel

$X_1 = \{o_j | Flu(o_j) = \{Yes\}\} = \{o_1, o_2, o_5, o_6\}$
$X_2 = \{o_j | Flu(o_j) = \{No\}\} = \{o_3, o_4\}$
$IND_C = \{\{o_1\}, \{o_2, o_3\}, \{o_4\}, \{o_5\}, \{o_6\}\}.$

B-Lower

$\underline{C}X_1 = \{o_1, o_5, o_6\}$
$\underline{C}X_2 = \{o_4\}$

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

B-Lower & B-Upper Approximation

Rappel

$X_1 = \{o_j | Flu(o_j) = \{Yes\}\} = \{o_1, o_2, o_5, o_6\}$
$X_2 = \{o_j | Flu(o_j) = \{No\}\} = \{o_3, o_4\}$
$IND_C = \{\{o_1\}, \{o_2, o_3\}, \{o_4\}, \{o_5\}, \{o_6\}\}.$

B-Lower

$\underline{C}X_1 = \{o_1, o_5, o_6\}$
$\underline{C}X_2 = \{o_4\}$

B-Lower

$\bar{C}X_1 = \{o_1, o_2, o_3, o_5, o_6\}$
$\bar{C}X_2 = \{o_2, o_3, o_4\}$

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Positive & Negative Region

Rappel

$\underline{C}X_1 = \{o_1, o_5, o_6\}$
$\underline{C}X_2 = \{o_4\}$
$\bar{C}X_1 = \{o_1, o_2, o_3, o_5, o_6\}$
$\bar{C}X_2 = \{o_2, o_3, o_4\}$

Positive & Negative Region

Rappel

$\underline{C}X_1 = \{o_1, o_5, o_6\}$
$\underline{C}X_2 = \{o_4\}$
$\bar{C}X_1 = \{o_1, o_2, o_3, o_5, o_6\}$
$\bar{C}X_2 = \{o_2, o_3, o_4\}$

Positive Region

$POS_C = \underline{C}X_1 \cup \underline{C}X_2$
$= \{o_1, o_5, o_6\} \cup \{o_4\}$
$= \{o_1, o_5, o_6, o_4\}$

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Positive & Negative Region

Rappel

$\underline{C}X_1 = \{o_1, o_5, o_6\}$
$\underline{C}X_2 = \{o_4\}$
$\bar{C}X_1 = \{o_1, o_2, o_3, o_5, o_6\}$
$\bar{C}X_2 = \{o_2, o_3, o_4\}$

Positive Region

$POS_C = \underline{C}X_1 \cup \underline{C}X_2$
$= \{o_1, o_5, o_6\} \cup \{o_4\}$
$= \{o_1, o_5, o_6, o_4\}$

Negative Region

$NEG_C = U - (\bar{C}X_1 \cup \bar{C}X_2)$
$= \{o_1, o_2, o_3, o_4, o_5, o_6\} - (\{o_1, o_2, o_3, o_5, o_6\} \cup \{o_2, o_3, o_4\})$
$= \emptyset$

Reduct

Définition

Un *reduct* est un sous ensemble minimal d'attributs ayant la même *Positive Region* que l'ensemble des attributs.

Reduct

Définition

Un *reduct* est un sous ensemble minimal d'attributs ayant la même *Positive Region* que l'ensemble des attributs.

Exemple

[["Headache", "Temperature"], ["Muscle-pain", "Temperature"]]

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Core

Définition

Un *core* est un ensemble d'attributs indépendant incluant tous les *reduct*.

Core

Définition

Un *core* est un ensemble d'attributs indépendant incluant tous les *reduct*.

Exemple

["Temperature"]

Dépendance des attributs

Rappel

$POS_C = \{o_1, o_5, o_6, o_4\}$
$U = \{o_1, o_2, o_3, o_4, o_5, o_6\}$

Exemple

$\lambda = POS_C / U$
$= |\{o_1, o_5, o_6, o_4\}| / |\{o_1, o_2, o_3, o_4, o_5, o_6\}|$
$= 4/6$

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

QuickReduct

- Nous calculons toutes les dépendances pour un attributs.

QuickReduct

- Nous calculons toutes les dépendances pour un attributs.
- Nous prenons l'attribut possèdent la dépendance la plus haute et nous l'ajoutons dans la réduction.

QuickReduct

- Nous calculons toutes les dépendances pour un attributs.
- Nous prenons l'attribut possèdent la dépendance la plus haute et nous l'ajoutons dans la réduction.
- Nous calculons la dépendance de la réduction en ajoutons à un à chaque attributs.

QuickReduct

- Nous calculons toutes les dépendances pour un attributs.
- Nous prenons l'attribut possèdent la dépendance la plus haute et nous l'ajoutons dans la réduction.
- Nous calculons la dépendance de la réduction en ajoutons à un à chaque attributs.
- Nous ajoutons l'attribut ayant la meilleur dépendance dans la réduction

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

QuickReduct

- Nous calculons toutes les dépendances pour un attributs.
- Nous prenons l'attribut possèdent la dépendance la plus haute et nous l'ajoutons dans la réduction.
- Nous calculons la dépendance de la réduction en ajoutons à un à chaque attributs.
- Nous ajoutons l'attribut ayant la meilleur dépendance dans la réduction
- Nous nous arrêtons quand nous obtenons la même dépendance qu'avec la totalité des attributs.

Application

Dataset *Statlog Heart*

	age	sex	chest pain type	blood pressure	cholestoral	blood sugar	electrocardiographic	heart rate max	angina	oldpeak	peak exercise	vessels	thal	heart disease
0	70.0	1.0	4.0	130.0	322.0	0.0	2.0	109.0	0.0	2.4	2.0	3.0	3.0	2
1	67.0	0.0	3.0	115.0	564.0	0.0	2.0	160.0	0.0	1.6	2.0	0.0	7.0	1
2	57.0	1.0	2.0	124.0	261.0	0.0	0.0	141.0	0.0	0.3	1.0	0.0	7.0	2
3	64.0	1.0	4.0	128.0	263.0	0.0	0.0	105.0	1.0	0.2	2.0	1.0	7.0	1
4	74.0	0.0	2.0	120.0	269.0	0.0	2.0	121.0	1.0	0.2	1.0	1.0	3.0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
265	52.0	1.0	3.0	172.0	199.0	1.0	0.0	162.0	0.0	0.5	1.0	0.0	7.0	1
266	44.0	1.0	2.0	120.0	263.0	0.0	0.0	173.0	0.0	0.0	1.0	0.0	7.0	1
267	56.0	0.0	2.0	140.0	294.0	0.0	2.0	153.0	0.0	1.3	2.0	0.0	3.0	1
268	57.0	1.0	4.0	140.0	192.0	0.0	0.0	148.0	0.0	0.4	2.0	0.0	6.0	1
269	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	2.0	3.0	3.0	2

270 rows × 14 columns

Figure – Extrait du dataset

Preprocessing

Preprocessing

- Détection des valeurs manquantes/ nulles.

Preprocessing

- Détection des valeurs manquantes/ nulles.
- Discretisation.

Attribut	Interval	Discretisation
age	[20 - 40)	Young
	[40 - 60)	Mid
	[60 - 80)	Old
Blood Pressure	[90 - 139]	Normal
	> 140	Abnormal
Cholestoral	[150 - 250]	Normal
	< 150 or > 250	Abnormal
Maximum heart rate	< 60	Low
	[60 - 100]	Normal
	> 100	Hight

Figure – Tableau de discrétisation

Preprocessing

- Détection des valeurs manquantes/ nulles.
- Discretisation.
- Encodage.

Preprocessing

- Détection des valeurs manquantes/ nulles.
- Discrétisation.
- Encodage.
- Équilibrer le nombre d'objet par classe.

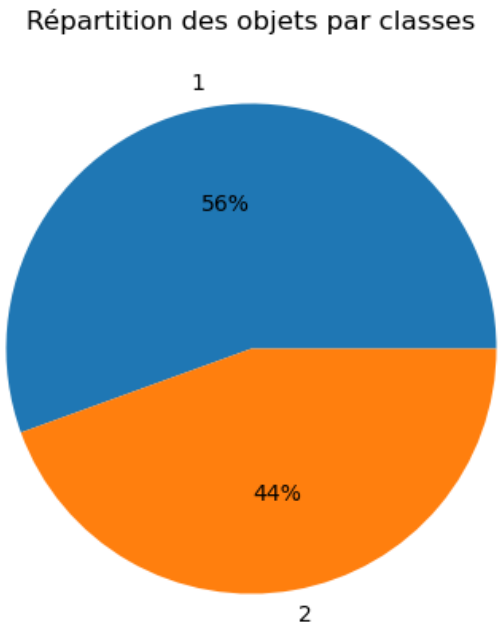


Figure – Répartition des objets par classe

Preprocessing

- Détection des valeurs manquantes/ nulles.
- Discrétisation.
- Encodage.
- Équilibrer le nombre d'objet par classe.
- Normalisation (attributs/ classes).

Data visualization

Data visualization

- Matrice de corrélation.

Data visualization

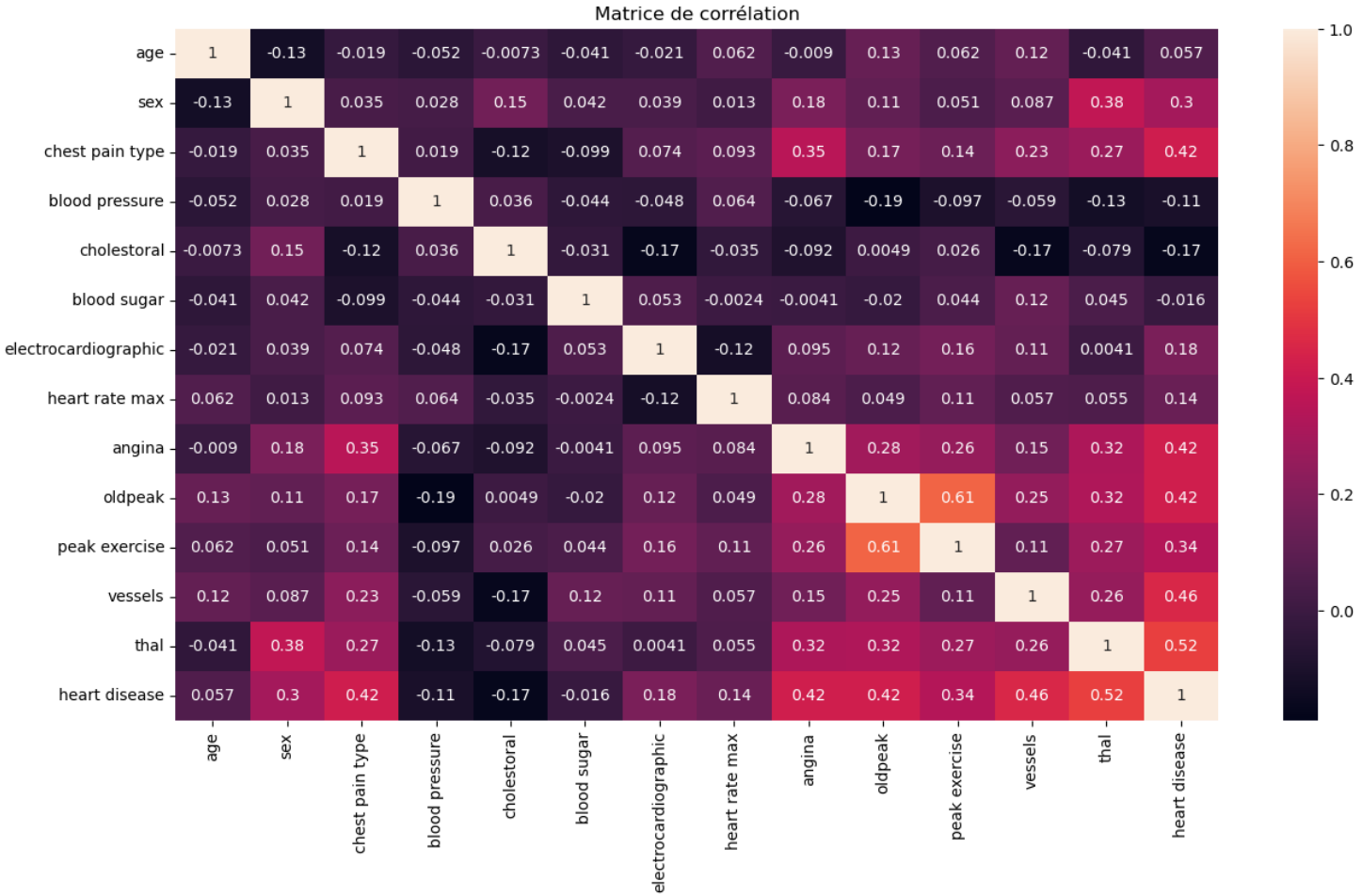


Figure – Matrice de corélation

Data visualization

- Matrice de corrélation.
- Boxplot => suppression "outliers".

Data visualization

- Matrice de corrélation.
- Boxplot => suppression "outliers".
- Histogrammes.

Data visualization

- Matrice de corrélation.
- Boxplot => suppression "outliers".
- Histogrammes.
- Pairplot.

Réductions trouvées

Liste des réductions trouvées

Méthode	age	sex	chest pain type	blood pressure	cholestoral	blood sugar	electrocardiographic	heart rate max	angina	oldpeak	peak exercise	vessels	thal
Reduct	✓	✓	✓	✓	X	X	✓	X	X	✓	X	✓	✓
QuickReduct	✓	✓	✓	✓	X	X	✓	X	X	✓	X	✓	✓
Variance	X	✓	X	✓	✓	X	✓	X	✓	X	X	X	✓
Fuzzy	✓	✓	✓	✓	✓	✓	✓	X	X	✓	X	X	X

Table – Liste des réductions

Liste des réductions trouvées

Méthode	age	sex	chest pain type	blood pressure	cholestoral	blood sugar	electrocardiographic	heart rate max	angina	oldpeak	peak exercise	vessels	thal
Reduct	✓	✓	✓	✓	X	X	✓	X	X	✓	X	✓	✓
QuickReduct	✓	✓	✓	✓	X	X	✓	X	X	✓	X	✓	✓
Variance	X	✓	X	✓	✓	X	✓	X	✓	X	X	X	✓
Fuzzy	✓	✓	✓	✓	✓	✓	✓	X	X	✓	X	X	X

Table – Liste des réductions

Liste des réductions trouvées

Méthode	age	sex	chest pain type	blood pressure	cholestoral	blood sugar	electrocardiographic	heart rate max	angina	oldpeak	peak exercise	vessels	thal
Reduct	✓	✓	✓	✓	X	X	✓	X	X	✓	X	✓	✓
QuickReduct	✓	✓	✓	✓	X	X	✓	X	X	✓	X	✓	✓
Variance	X	✓	X	✓	✓	X	✓	X	✓	X	X	X	✓
Fuzzy	✓	✓	✓	✓	✓	✓	✓	X	X	✓	X	X	X

Table – Liste des réductions

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Liste des réductions trouvées

Méthode	age	sex	chest pain type	blood pressure	cholestoral	blood sugar	electrocardiographic	heart rate max	angina	oldpeak	peak exercise	vessels	thal
Reduct	✓	✓	✓	✓	X	X	✓	X	X	✓	X	✓	✓
QuickReduct	✓	✓	✓	✓	X	X	✓	X	X	✓	X	✓	✓
Variance	X	✓	X	✓	✓	X	✓	X	✓	X	X	X	✓
Fuzzy	✓	✓	✓	✓	✓	✓	✓	X	X	✓	X	X	X

Table – Liste des réductions

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Liste des réductions trouvées

Méthode	age	sex	chest pain type	blood pressure	cholestoral	blood sugar	electrocardiographic	heart rate max	angina	oldpeak	peak exercise	vessels	thal
Reduct	✓	✓	✓	✓	X	X	✓	X	X	✓	X	✓	✓
QuickReduct	✓	✓	✓	✓	X	X	✓	X	X	✓	X	✓	✓
Variance	X	✓	X	✓	✓	X	✓	X	✓	X	X	X	✓
Fuzzy	✓	✓	✓	✓	✓	✓	✓	X	X	✓	X	X	X

Table – Liste des réductions

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Liste des réductions trouvées

Méthode	age	sex	chest pain type	blood pressure	cholestoral	blood sugar	electrocardiographic	heart rate max	angina	oldpeak	peak exercise	vessels	thal
Reduct	✓	✓	✓	✓	X	X	✓	X	X	✓	X	✓	✓
QuickReduct	✓	✓	✓	✓	X	X	✓	X	X	✓	X	✓	✓
Variance	X	✓	X	✓	✓	X	✓	X	✓	X	X	X	✓
Fuzzy	✓	✓	✓	✓	✓	✓	✓	X	X	✓	X	X	X

Table – Liste des réductions

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Classification

Modèles de classification

Liste des modèles appliqués :

- Gaussian Process

Modèles de classification

Liste des modèles appliqués :

- Gaussian Process
- Random Forest

Modèles de classification

Liste des modèles appliqués :

- Gaussian Process
- Random Forest
- Nearest Neighbors

Modèles de classification

Liste des modèles appliqués :

- Gaussian Process
- Random Forest
- Nearest Neighbors
- Autres (Linear SVM, RBF SVM, Neural Net, Naive Bayes, Decision Tree, AdaBoost, QuadraticDiscriminantAnalysis)

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES

Comparaison

Matrice de corrélation

```
sans_reduc

[array([[23,  4],
        [ 5, 15]], dtype=int64),
 array([[22,  5],
        [ 5, 15]], dtype=int64),
 array([[24,  3],
        [ 8, 12]], dtype=int64)]
```

Figure – Matrice de confusion de sans *feature selection*

Matrice de corrélation

```
reduct
Gaussian Process
[[-1  1]
 [-5  5]]
Random Forest
[[ 2 -2]
 [-4  4]]
Nearest Neighbors
[[-2  2]
 [-3  3]]
```

Figure – Différence entre avec *reduct* et sans

Matrice de corrélation

```
quickreduct
Gaussian Process
[[-1  1]
 [-5  5]]
Random Forest
[[ 2 -2]
 [-3  3]]
Nearest Neighbors
[[-2  2]
 [-3  3]]
```

Figure – Différence entre avec *quickreduct* et sans

Matrice de corrélation

```
variance
Gaussian Process
[[-4  4]
 [-4  4]]
Random Forest
[[-3  3]
 [-3  3]]
Nearest Neighbors
[[-3  3]
 [-2  2]]
```

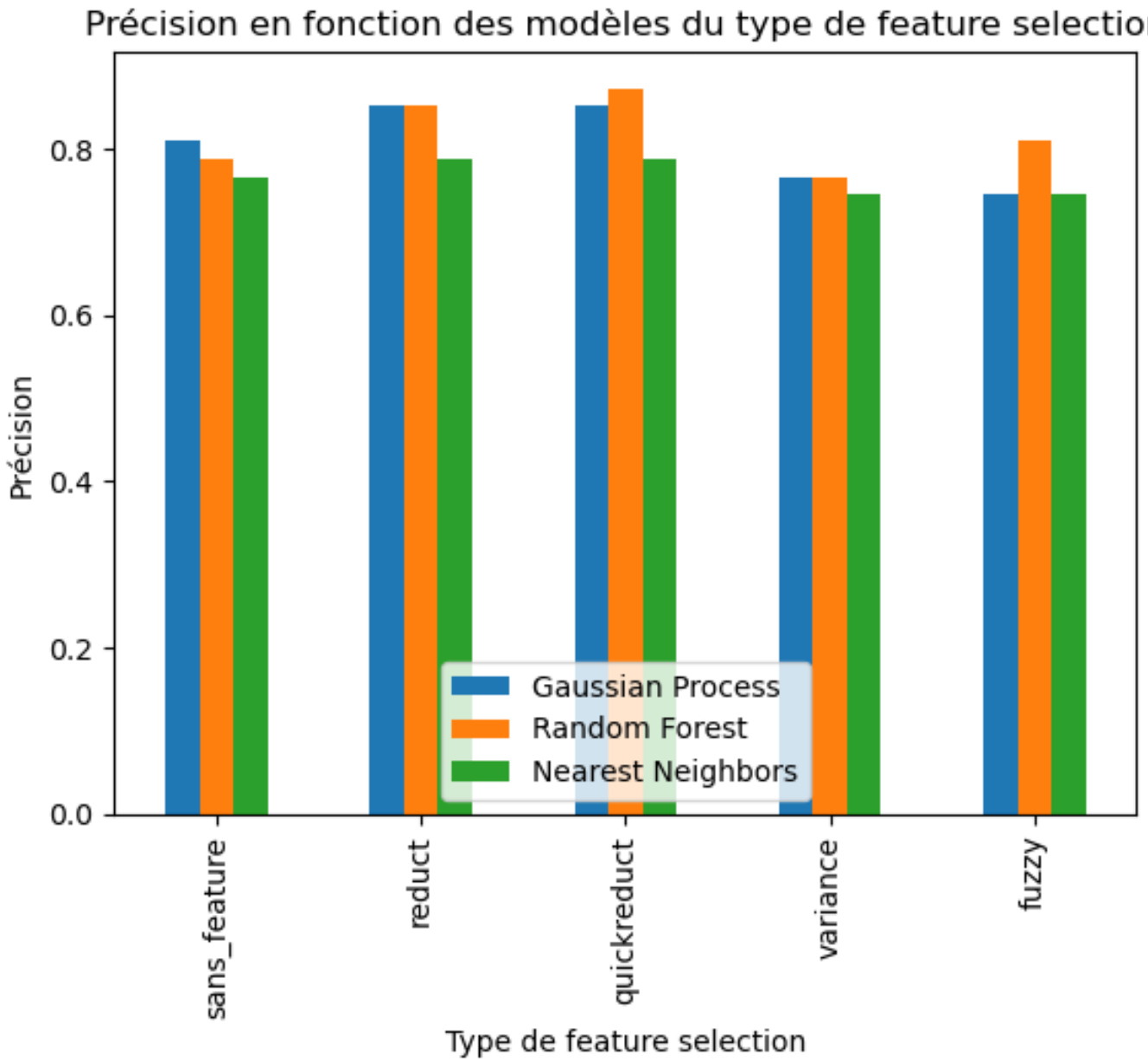
Figure – Différence entre avec *variance* et sans

Matrice de corrélation

```
fuzzy
Gaussian Process
[[-2  2]
 [-1  1]]
Random Forest
[[ 0  0]
 [-1  1]]
Nearest Neighbors
[[-2  2]
 [-1  1]]
```

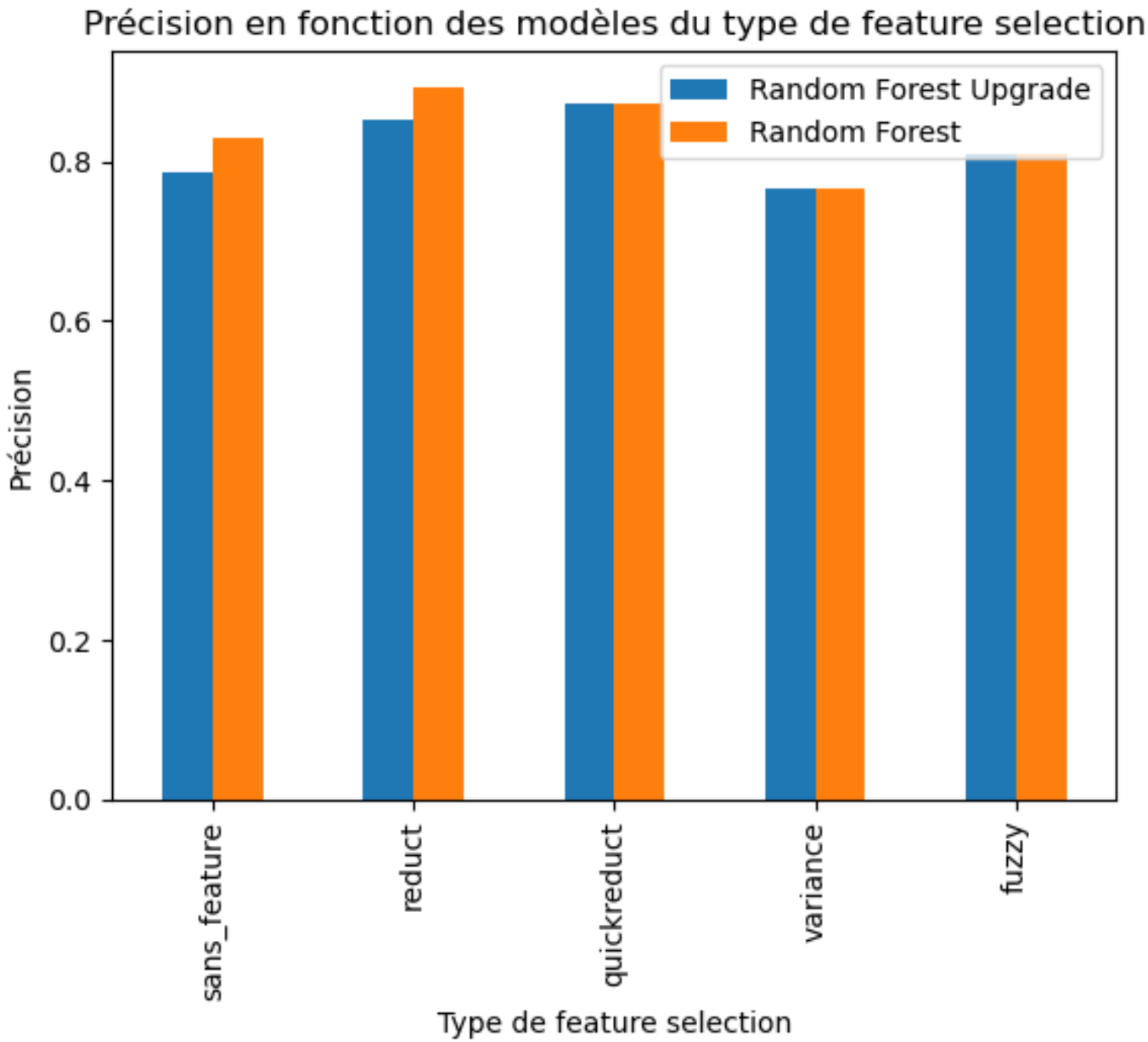
Figure – Différence entre avec *fuzzy* et sans

Comparaison de la précision



Amélioration des modèles.

Amélioration du *Random Forest*



Limite du Quickreduct

Limite

- Générer tous les sous ensembles possibles d'attributs.

Limite

- Générer tous les sous ensembles possibles d'attributs.
- Re-calcul tout à chaque fois.

Amélioration du Quickreduct

Améliorations

- Stocker les résultats et faire une version incrémentale.

Améliorations

- Stocker les résultats et faire une version incrémentale.
- Trouver un moyen de partir de l'ensemble des attributs et de supprimer les moyens utiles si c'est plus rapide (comment le savoir ?).

Références

Oracle.
Qu'est-ce que le data mining ?

Merci pour votre attention !
Avez vous des questions ?

UVSQ université PARIS-SACLAY UFR des Sciences CAMPUS DE VERSAILLES