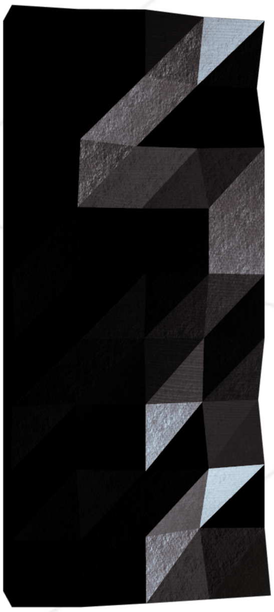


Training Piscine datascience - 3

TO DO

Summary: Today, you will see some TODO.

Version: 1.00

Contents

Chapter I

General rules

- Your project must be realized in a virtual machine.
- Your virtual machine must have all the necessary software to complete your project. These softwares must be configured and installed.
- You can choose the operating system to use for your virtual machine.
- You must be able to use your virtual machine from a cluster computer.
- You must use a shared folder between your virtual machine and your host machine.
- During your evaluations you will use this folder to share with your repository.
- Your functions should not quit unexpectedly (segmentation fault, bus error, double free, etc) apart from undefined behaviors. If this happens, your project will be considered non functional and will receive a `0` during the evaluation.
- We encourage you to create test programs for your project even though this work **won't have to be submitted and won't be graded**. It will give you a chance to easily test your work and your peers' work. You will find those tests especially useful during your defence. Indeed, during defence, you are free to use your tests and/or the tests of the peer you are evaluating.
- Submit your work to your assigned git repository. Only the work in the git repository will be graded. If Deepthought is assigned to grade your work, it will be done after your peer-evaluations. If an error happens in any section of your work during Deepthought's grading, the evaluation will stop.

Chapter II

Specific instructions of the day

Module de data Scientist, ils utilisent souvent comme techno le python, Jupyter Notebook ...

A vous de trouver les outils qui vous conviennent ce module est en language libre.

Le role du data Scientist est de predire "l'avenir" avec des modele d'aprentissage automatique sur des donnes passees, il doit etre force de proposition pour expliquer l'interet possitif a la mise en place de ses modeles, creer des outils d'aide a la prise de decision

Chapter III

Exercise 00

	Exercise 00
Exercice 00 : TO DO	
Turn-in directory : *ex*00/	
Files to turn in : `TO DO.*`	
Allowed functions : `None`	

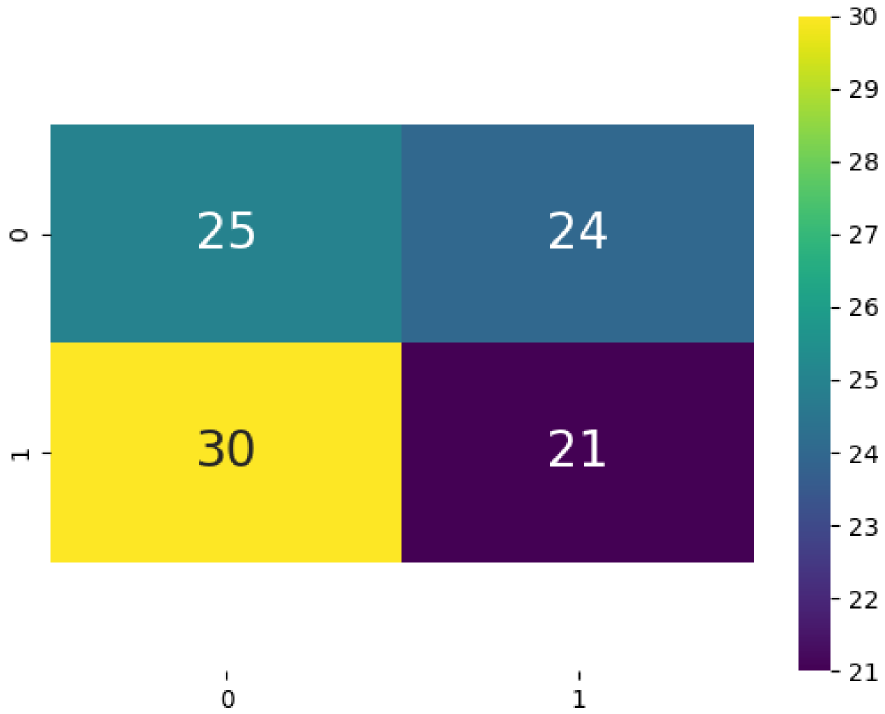
- faire un prog qui prend en arg 1er fichier avec les predictions, 2nd fichier les truth, puis print
- faire une Confusion matrix

```
            precision   recall f1-score    total

         0      0.51     0.45     0.48       55
         1      0.41     0.47     0.44       45

         accuracy                 0.46      100

[[25 24]
 [30 21]]
```



Si cette exo est faux, on ne passe pas a la suite, c'est finito

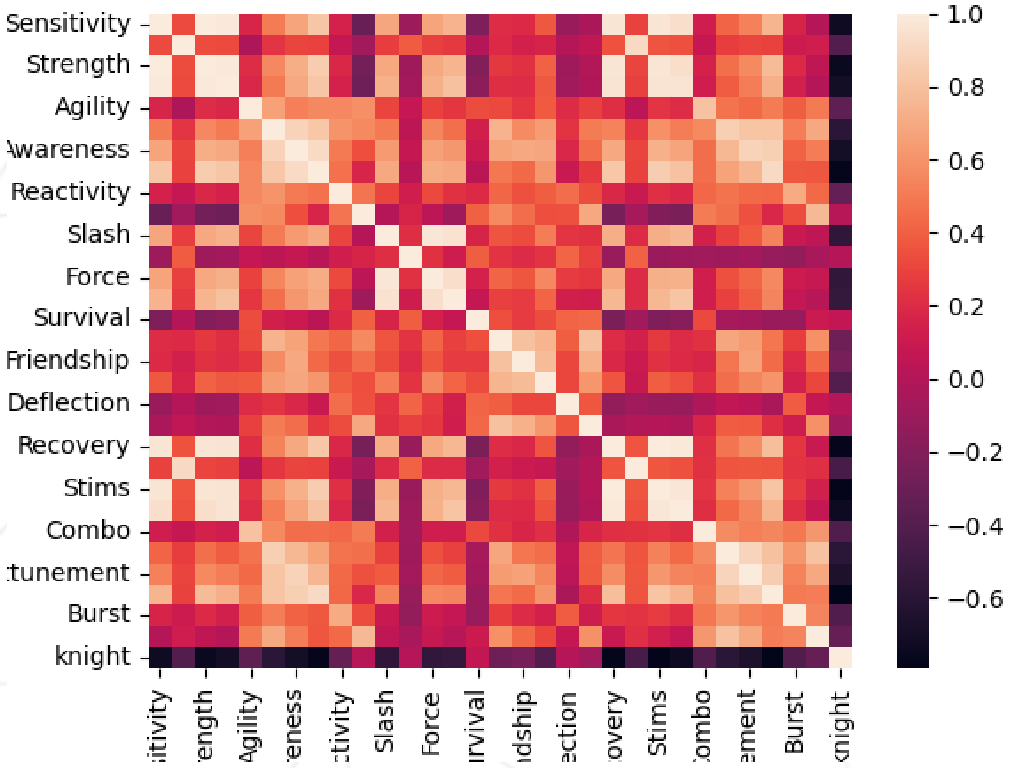
soyez sur d'avoir bien compris cette exo car cela va etre verifie

Chapter IV

Exercise 01

	Exercise 01
Exercice 01 :	
Turn-in directory : *ex*01/	
Files to turn in : `TO DO.*`	
Allowed functions : `None`	

- Faire une Heatmap pour voir le Correlation Coefficient entre les donees

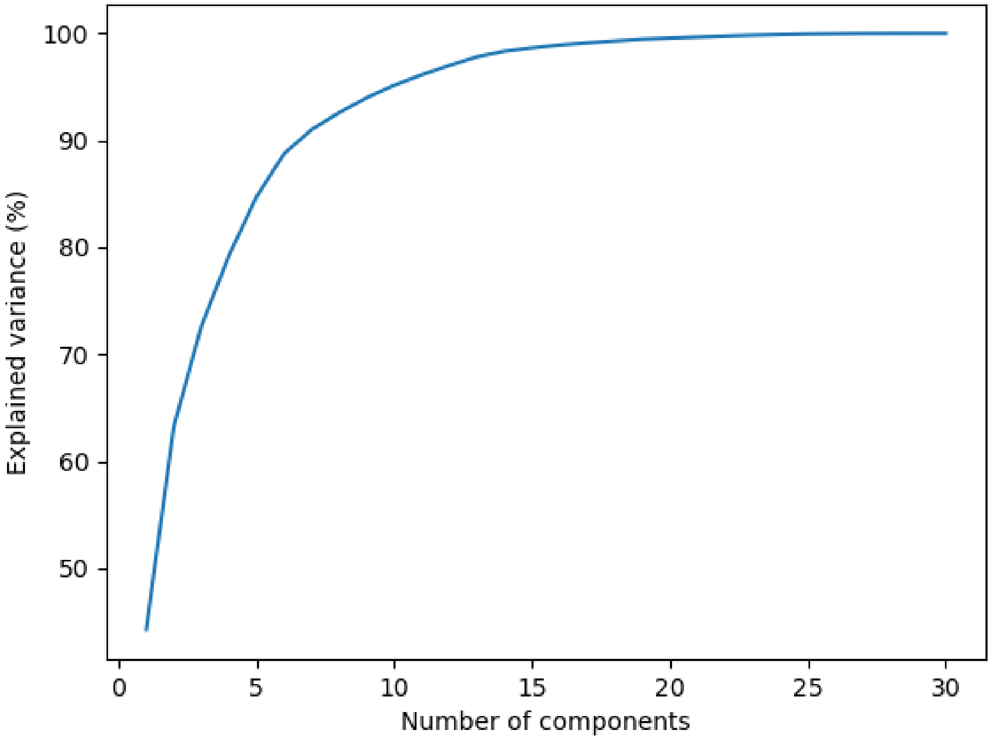


Chapter V

Exercise 02

	Exercise 02
Exercice 02 :	
Turn-in directory : *ex*02/	
Files to turn in : `TO DO.*`	
Allowed functions : `None`	

- additioner les variances pour voir combien sont utiles dans 90% du modele



Chapter VI

Exercise 03

	Exercise 03
Exercice 03 : Feature Selection	
Turn-in directory : *ex*03/	
Files to turn in : `TO DO.*`	
Allowed functions : `None`	

les donnee semble trop Multicollinearity, vous aller devoir faire un modele de Detecting Multicollinearity

Il en existe un certain nombre de modele de detection d'importance de variables (Lasso, Backward Elimination, Step Forward Selection, ...)

Mais ici vous aller devoir utiliser le Variance Inflation Factor (VIF)

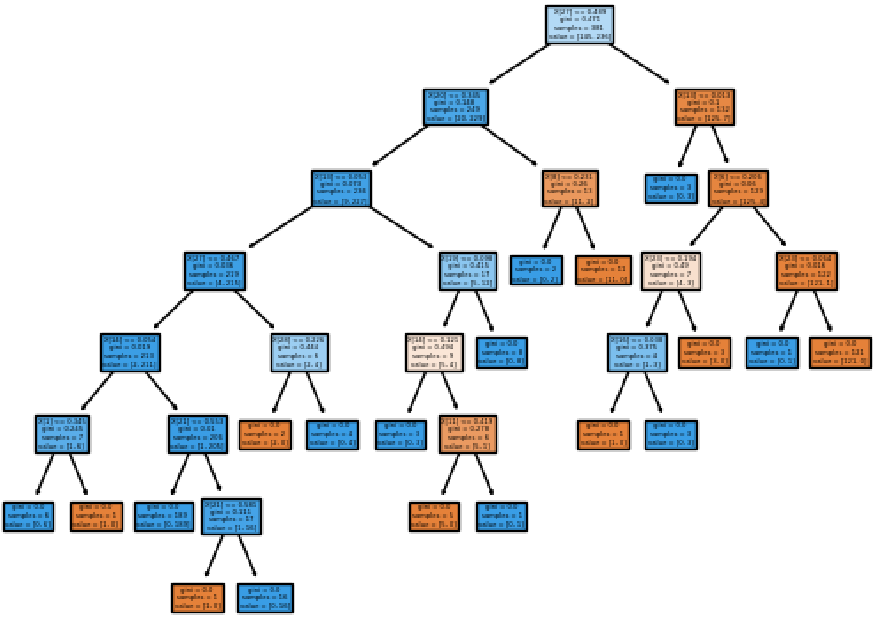
- Le VIF de vos features doit etre inferieur a 5
- pas de hard coding (on change les data dans l'eval)

Chapter VII

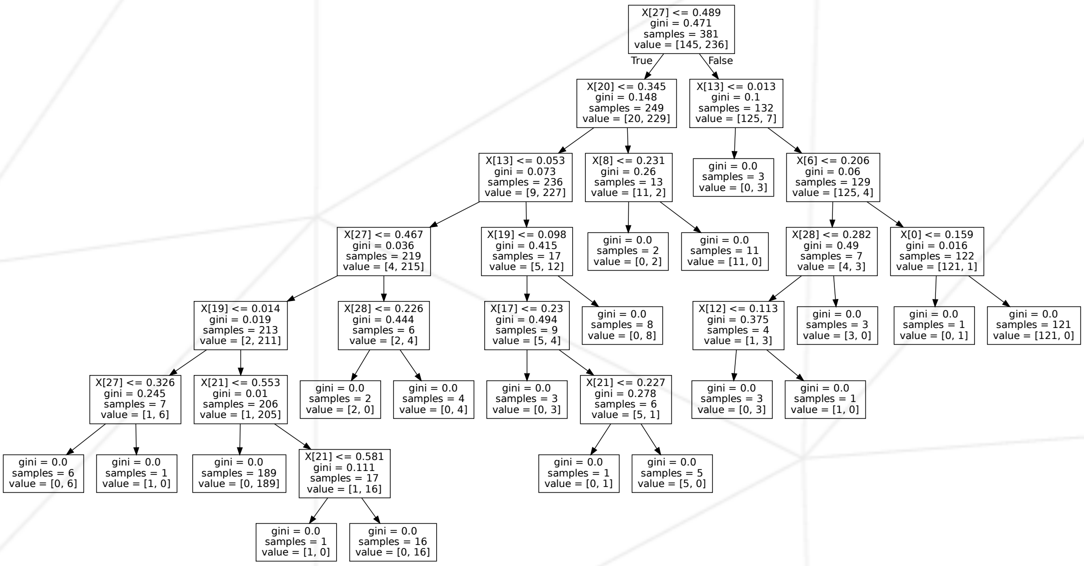
Exercise 04

	Exercise 04
Exercice 04 : TO DO	
Turn-in directory : *ex*04/	
Files to turn in : `TO DO.*`	
Allowed functions : `None`	

- faire un modele de Decision Tree Classifier ou Random Forest Classifier (l'aisser de choix pour que le stud cherche ?)
- faire un prog qui prend en arg le fichier de train en 1er arg et le fichier de test en 2nd arg et qui ecrit un fichier avec les reponses dedans (comme dans dslr)
- afficher l'arbre dans un graphique



Decision tree trained on all Knights features

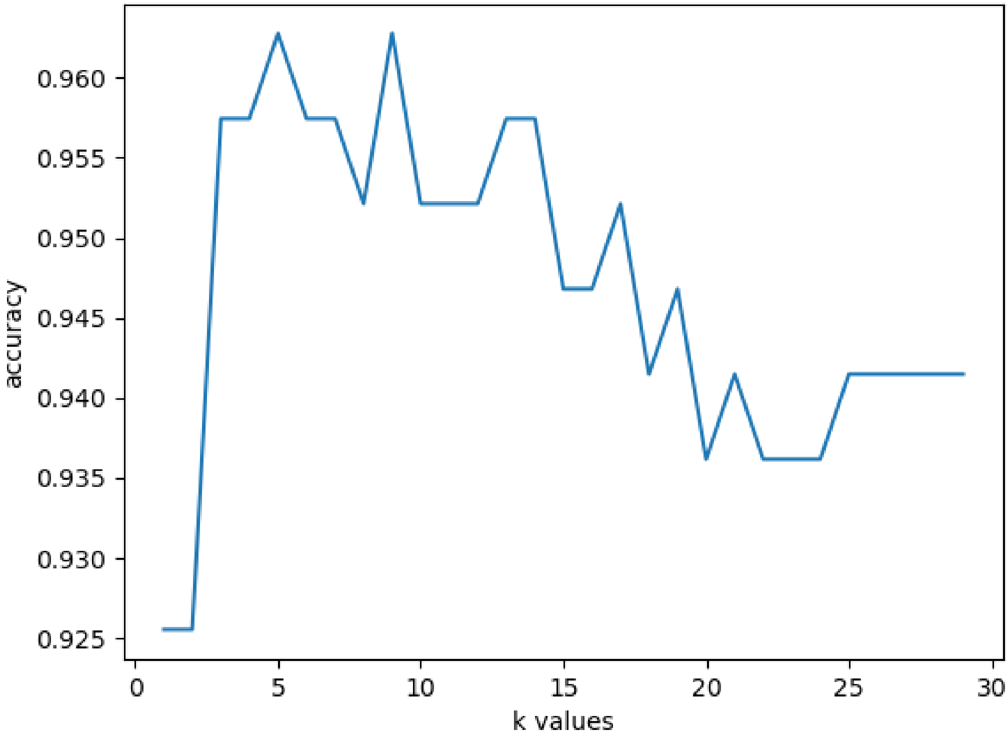


Chapter VIII

Exercise 05

	Exercise 05
Exercice 05 : TO DO	
Turn-in directory : *ex*05/	
Files to turn in : `TO DO.*`	
Allowed functions : `None`	

- faire un KNN qui calcul le % de precision en fonction du nb de k value
- afficher le graph
- faire un prog qui prend en arg le fichier de train en 1er arg et le fichier de test en 2nd arg et qui ecrit un fichier avec les reponses dedans (comme dans dslr)


0.960
0.955
0.950
0.945
0.940
0.935
0.930
0.925
accuracy
0
5
10
15
20
25
30
k values

Chapter IX

Submission and peer-evaluation

Turn in your assignment in your `Git` repository as usual. Only the work inside your repository will be evaluated during the defense. Don't hesitate to double check the names of your folders and files to ensure they are correct.

> The evaluation process will happen on the computer of the evaluated group.