Rapport Projet Machine Learning

Il s'agit d'un rapport dans le cadre du projet de la matière Machine Learning, formation MIAGE M1 en apprentissage, université PSL - Paris Dauphine, année scolaire 2023 - 2024. Le projet est pour but d'implémenter la régression logistique et mettre en œuvre une méthodologie solide.

Étudiant: Binh Minh NGUYEN

Sommaire

Sommaire
Question I.6
Question II.7
Question II.9
Question III.10
- Contexte
- Origine
- Nombre d'observations
- Variables
- Classification
- Instructions d'utilisation

Question III.11
Question III.12
- L'approche choisie: Arbres de décisions
- Comparer cette 3ème approche avec les 2 autres

Annexe: Table comparative complète des 3 approches pour 4 datasets

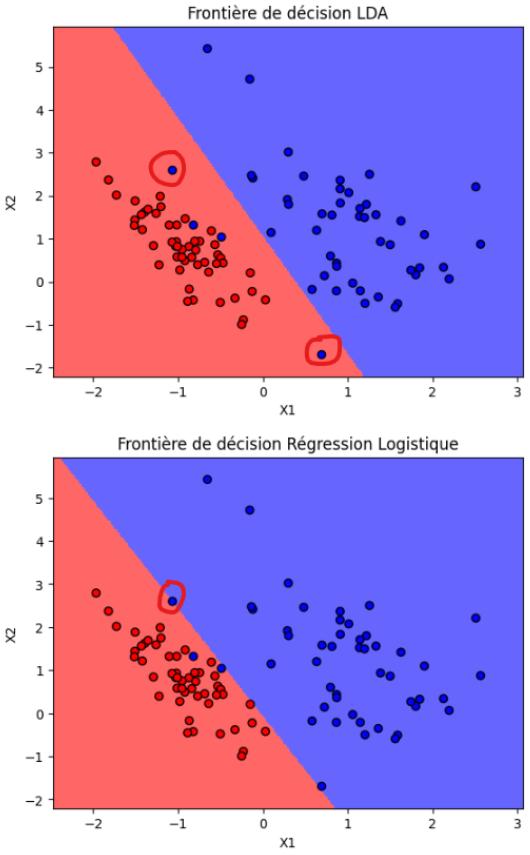
Question I.6

Sur les données d'entraînement, la régression logistique présente une erreur légèrement inférieure à celle de LDA (0.025 contre 0.0375). Cependant, sur les données de test, les deux modèles affichent une erreur identique (0.05).

```
beta= [[7.09913719]
 [2.83601857]
 [0.19272345]]
Logistic regression says that x belongs to class 1
Error ratio of model generated by Logistic Regression on training data:  0.025
Error ratio of model generated by Logistic Regression on testing data:  0.05
w= [-7.30352402 -2.66269449] , b= 2.842390511575443
Error ratio of model generated by LDA on training data:  0.0375
Error ratio of model generated by LDA on testing data:  0.05
Error ratio of model generated by Logistic Regression on training data:  0.025
Error ratio of model generated by Logistic Regression on testing data:  0.05
Logistic regression says that x belongs to class 1
LDA says that x belongs to class 0
```

Une divergence est observée dans les prédictions des deux modèles. La régression logistique prédit la classe 1 pour x, tandis que LDA prédit la classe 0.

La frontière de décision de la régression logistique est tracée de manière plus précise, car elle optimise la couverture maximale des points bleus dans la zone bleue, tout en minimisant la largeur de la zone rouge nécessaire pour englober l'ensemble des points rouges. Comme illustré, dans le cas de LDA, la zone bleue ne peut pas inclure les deux points encerclés. En revanche, dans le cas de la régression logistique, la zone bleue peut toujours englober ces deux points, tout en évitant toute inclusion erronée de points rouges supplémentaires.



Ainsi, la différence entre ces deux frontières de décision correspond effectivement à la disparité des taux d'erreur entre les deux modèles.

Cependant, la pente de la frontière de décision semble être la même dans les deux cas.

Question II.7

Après avoir inclus une ligne permettant d'ajouter une valeur aberrante dans les données d'apprentissage, les paramètres (Beta, w et b) des deux modèles ont subi des changements significatifs. Le taux d'erreur des deux modèles sur les données d'entraînement est légèrement plus élevé (0,125), mais reste inchangé sur les données de test. Le modèle de régression logistique prédisait initialement que x

appartiendrait à la classe 1, mais prédit désormais la classe 0. En revanche, le modèle de LDA maintient sa prédiction.

Avant

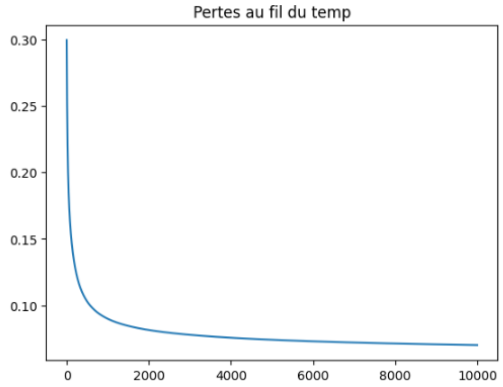
```
attain itermax
beta= [[7.08312611]
 [2.82794967]
 [0.18958058]]
w= [-7.30352402 -2.66269449] , b= 2.842390511575443
Error ratio of model generated by LDA on training data:  0.0375
Error ratio of model generated by LDA on testing data:  0.05
Error ratio of model generated by Logistic Regression on training data:  0.025
Error ratio of model generated by Logistic Regression on testing data:  0.05
Logistic regression says that x belongs to class 1
LDA says that x belongs to class 0
```

Après

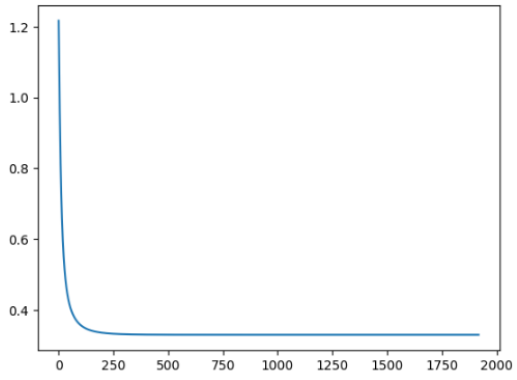
```
rising loss
new beta= [[ 2.19516933]
 [ 0.5047935 ]
 [-0.355884  ]]
new w= [-2.17522915 -0.32527617] , new b= 0.47301179099839974
Error ratio of new model generated by LDA on new training data:  0.05
Error ratio of new model generated by LDA on testing data:  0.05
Error ratio of new model generated by Logistic Regression on new training data:  0.0375
Error ratio of new model generated by Logistic Regression on new testing data:  0.05
Logistic regression says that x belongs to class 0
LDA says that x belongs to class 0
```

Auparavant, l'algorithme de régression logistique s'arrêtait lorsqu'il atteignait le nombre maximal d'itérations, mais maintenant il se termine après environ 2000 itérations, avec la raison d'arrêt indiquée comme "une nouvelle perte plus grande que l'ancienne" (l'itérateur commence à “descendre plutôt qu'à monter sur le courbe”).

Avant



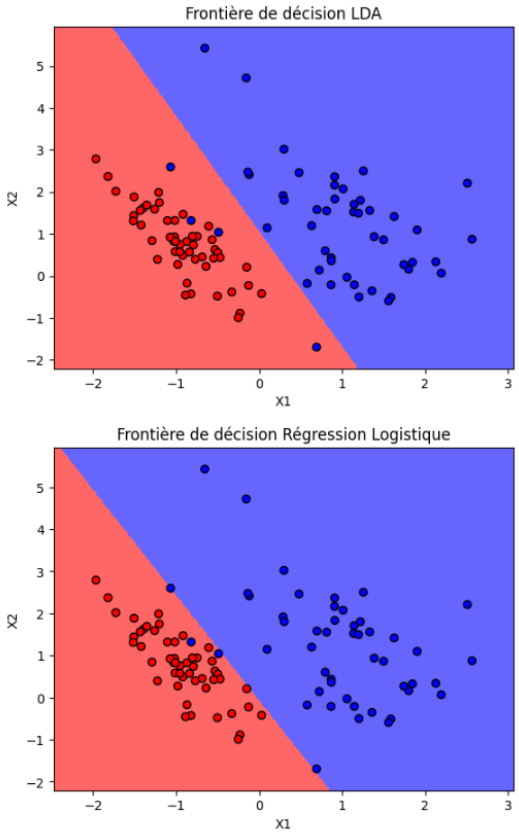
Après



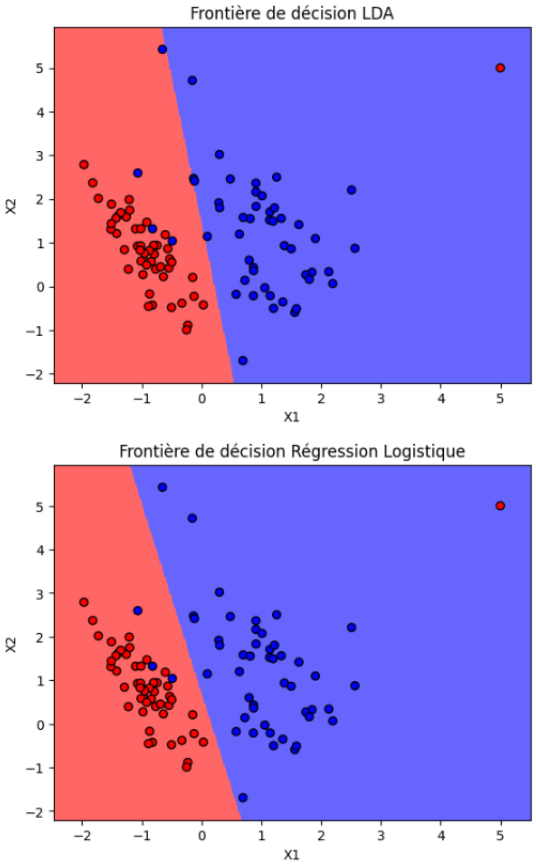
Comme les paramètres (Beta, w et b) des deux modèles, les frontières de décision ont également subi d'importants changements. Cette fois-ci, les zones de couleur sont encore plus précisément délimitées dans le cas de la régression logistique. Toutefois, les pentes des deux frontières semblent légèrement différentes cette fois-ci, la pente dans le cas de la régression logistique semblant être un peu plus raide.

De plus, après l'ajout d'un point aberrant, la pente de la frontière de décision dans le cas A change peu par rapport à la pente de la frontière de décision dans le cas B. Ceci est raisonnable car la régression logistique est souvent plus robuste aux valeurs aberrantes et aux distributions non gaussiennes.

Avant



Après



Question II.9

Pour cette question, j'ai augmenté la dimension des données à 7, dont 2 variables informatives, 2 redondantes et le reste sont des variables inutiles.

```
beta= [[ 2.38859968]
 [ 6.75679694]
 [-0.99414543]
 [-1.42175641]
 [-2.71708737]
 [-3.83269462]
 [ 0.52182766]
 [-0.36056053]]
attain itermax
w= [-2.5        -9.875       0.59641513  0.54642974 -1.          0.5
 -0.40696227] , b= 3.7174999152419015
Error ratio of new model generated by LDA on training data:  0.0375
Error ratio of new model generated by LDA on testing data:  0.15
Error ratio of new model generated by Logistic Regression on training data:  0.0125
Error ratio of new model generated by Logistic Regression on testing data:  0.1
Mean errors Logistic Regression 0.07
Var errors Logistic Regression 0.0651
Mean errors LDA 0.1276595744680851
Var errors LDA 0.11136260751471258
```

Cette fois, le modèle de régression logistique comporte nettement moins d'erreurs que le modèle B (0,0125 contre 0,0375 sur les données d'entraînement, 0,1 contre 0,15 sur les données de test). Lors de la validation croisée à n blocs, on observe que la moyenne et la variance des taux d'erreur du modèle de régression logistique sont également inférieures à celles de LDA (Moyenne : 0,07 contre 0,1277 ; Variance : 0,0651 contre 0,1114). Cela peut être dû au fait que lorsque le nombre de dimensions est élevé (7), le modèle LDA doit estimer beaucoup plus de paramètres que le modèle de régression logistique (35 contre 8), car il nécessite l'estimation de 7*(7-1)/2+7*2=35 paramètres, tandis que la régression logistique en nécessite seulement 7+1=8.

Question III.10

Contexte

Crise cardiaque

Une crise cardiaque (maladies cardiovasculaires) survient lorsque le flux sanguin vers le muscle cardiaque est soudainement bloqué. Selon les statistiques de l'OMS, chaque année, 17,9 millions de personnes décèdent des suites d'une crise cardiaque, faisant de cette affection un problème de santé mondial majeur. Cette réalité nécessite une compréhension approfondie de ses précurseurs et des facteurs atténuants potentiels.

Les études médicales indiquent que le mode de vie humain est la principale cause de ce problème cardiaque. Outre cela, de nombreux facteurs clés avertissent qu'une personne peut présenter un risque de crise cardiaque. Cet ensemble de données englobe un large éventail d'attributs, tels que l'âge, le sexe, le taux de cholestérol, la tension artérielle, etc., dans le but d'élucider l'interaction complexe de ces variables pour déterminer la probabilité d'une crise cardiaque.

En utilisant l'analyse prédictive et l'apprentissage automatique sur cet ensemble de données, les chercheurs et les professionnels de la santé peuvent parvenir à une fonctionnalité de classification binaire cruciale indiquant la présence ou l'absence d'un risque de crise cardiaque. Cette approche permet de développer des stratégies proactives de prévention et de gestion des maladies cardiaques. L'ensemble de données témoigne des efforts collectifs déployés pour améliorer notre compréhension de la santé cardiovasculaire et ouvrir la voie à un avenir plus sain.

Origine

Prit Sheta, 2021, "Heart Attack," Kaggle, https://www.kaggle.com/datasets/pritsheta/heart-attack/data

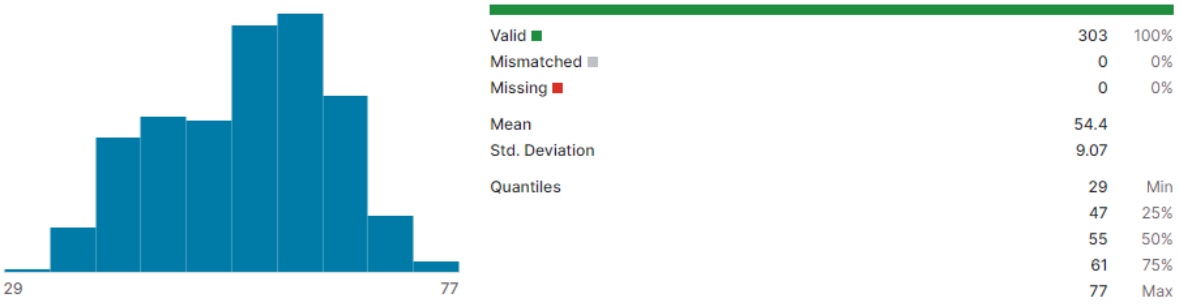
Nombre d'observations

303

Variables

13 variables de types d’entier et de flottant

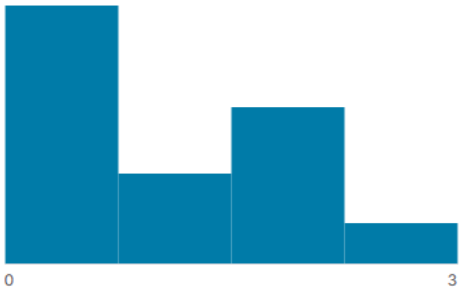
- **age** : Âge en années



- **sex**: Sexe:
  - 1 = masculin
  - 0 = féminin

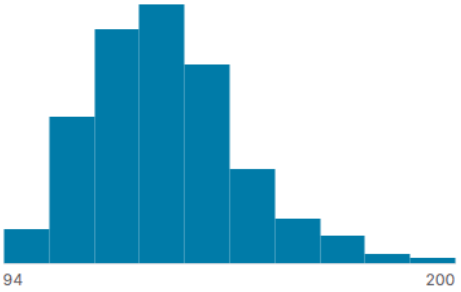
Valid	303	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.68	
Std. Deviation	0.47	
Quantiles	0	Min
	0	25%
	1	50%
	1	75%
	1	Max

- **cp**: Type péricardite constrictive:
  - 1 = angine typique (tous les critères présents)
  - 2 = angine atypique (deux des trois critères satisfaits)
  - 3 = douleur non angineuse (moins d'un critère satisfait)
  - 4 = asymptomatique (aucun des critères n'est satisfait)



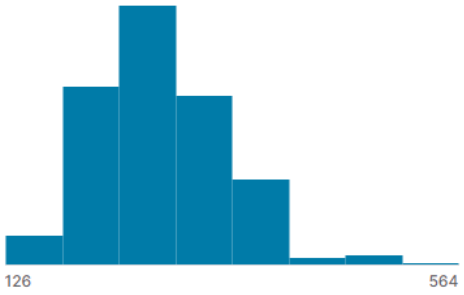
Valid	303	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.97	
Std. Deviation	1.03	
Quantiles	0	Min
	0	25%
	1	50%
	2	75%
	3	Max

- **trestbps**: Tension artérielle au repos (en mmHg, à l'admission à l'hôpital)



Valid	303	100%
Mismatched	0	0%
Missing	0	0%
Mean	132	
Std. Deviation	17.5	
Quantiles	94	Min
	120	25%
	130	50%
	140	75%
	200	Max

- **chol**: Cholestérol sérique en mg/dL



Valid	303	100%
Mismatched	0	0%
Missing	0	0%
Mean	246	
Std. Deviation	51.7	
Quantiles	126	Min
	211	25%
	240	50%
	275	75%
	564	Max

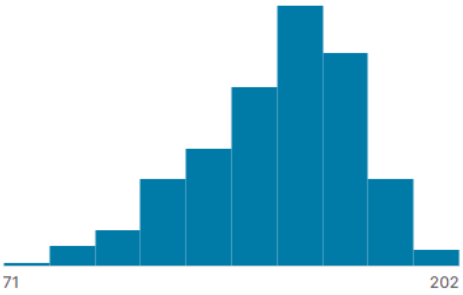
- **fbs**: Glycémie à jeun > 120 mg/dL (susceptible d'être diabétique):
  - 1 = vrai
  - 0 = faux

Valid	303	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.15	
Std. Deviation	0.36	
Quantiles	0	Min
	0	25%
	0	50%
	0	75%
	1	Max

- **restecg**: Résultats de l'électrocardiogramme de repos:
  - 0 = normal
  - 1 = présentant une anomalie de l'onde ST-T (inversions de l'onde T et/ou élévation ou dépression ST > 0,05 mV)
  - 2 = montrant une anomalie ventriculaire gauche probable ou certaine hypertrophie selon les critères d'Estes

Valid	303	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.53	
Std. Deviation	0.52	
Quantiles	0	Min
	0	25%
	1	50%
	1	75%
	2	Max

- **thalach**: Le plus grand nombre de battements par minute que votre cœur peut atteindre lors d'un exercice intense et intense.

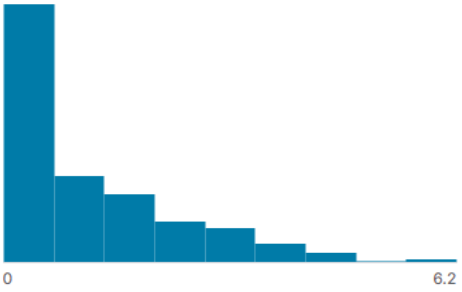


Valid	303	100%
Mismatched	0	0%
Missing	0	0%
Mean	150	
Std. Deviation	22.9	
Quantiles	71	Min
	133	25%
	153	50%
	166	75%
	202	Max

- **exang**: Angine induite par l'effort:
  - 1 = oui
  - 0 = non

Valid	303	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.33	
Std. Deviation	0.47	
Quantiles	0	Min
	0	25%
	0	50%
	1	75%
	1	Max

- **oldpeak**: Dépression ST induite par l'exercice par rapport au repos (en mm, obtenue en soustrayant les points du segment ST les plus bas pendant l'exercice et le repos)

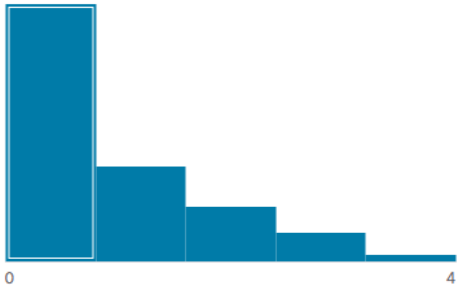


Valid	303	100%
Mismatched	0	0%
Missing	0	0%
Mean	1.04	
Std. Deviation	1.16	
Quantiles	0	Min
	0	25%
	0.8	50%
	1.6	75%
	6.2	Max

- **slope**: La pente du segment ST d'effort maximal, les anomalies ST-T sont considérées comme un indicateur crucial pour identifier la présence d'ischémie:
  - 1 = ascendante
  - 2 = plate
  - 3 = descendante

Valid	303	100%
Mismatched	0	0%
Missing	0	0%
Mean	1.4	
Std. Deviation	0.62	
Quantiles	0	Min
	1	25%
	1	50%
	2	75%
	2	Max

- **ca**: Nombre de vaisseaux majeurs (0-3) colorés par fluoroscopie. Les principaux vaisseaux cardiaques sont les suivants : l'aorte, la veine cave supérieure, la veine cave inférieure, l'artère pulmonaire (sang pauvre en oxygène → poumons), les veines pulmonaires (sang riche en oxygène → cœur) et les artères coronaires (qui alimentent en sang tissu cardiaque).



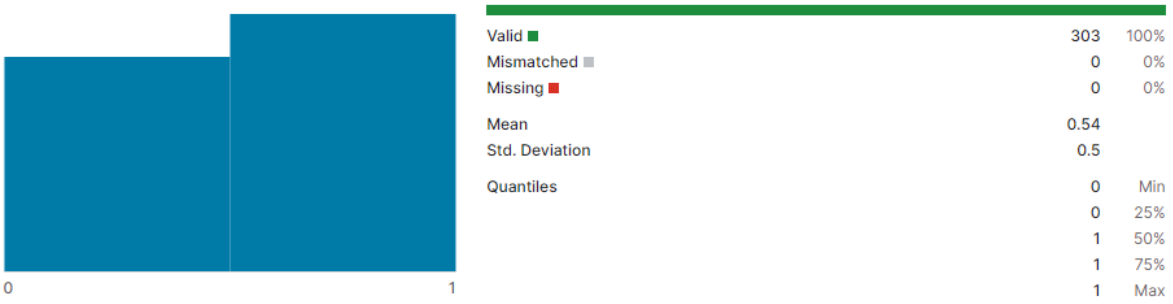
Valid	303	100%
Mismatched	0	0%
Missing	0	0%
Mean	0.73	
Std. Deviation	1.02	
Quantiles	0	Min
	0	25%
	0	50%
	1	75%
	4	Max

- **thal**: Thallium:
  - 0 = normal
  - 1 = défaut corrigé (le tissu cardiaque ne peut pas absorber le thallium à la fois sous stress et au repos)
  - 2 = défaut réversible (le tissu cardiaque est incapable d'absorber le thallium uniquement pendant la partie exercice du test)

Valid	303	100%
Mismatched	0	0%
Missing	0	0%
Mean	2.31	
Std. Deviation	0.61	
Quantiles	0	Min
	2	25%
	2	50%
	3	75%
	3	Max

Classification

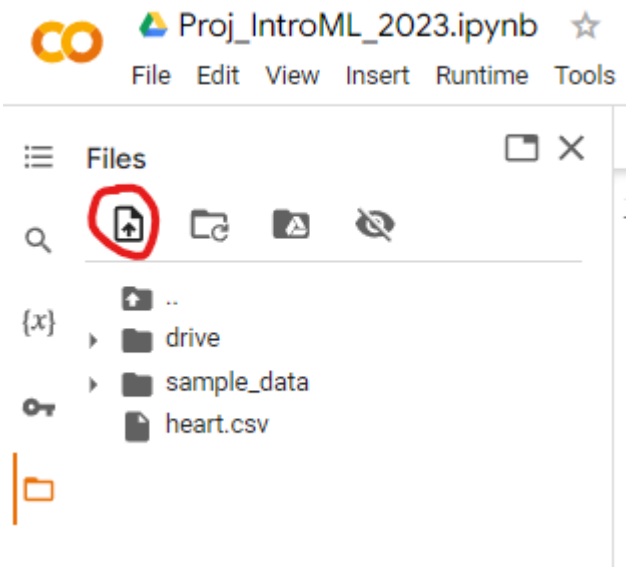
- 0 = absence d'un risque de crise cardiaque
- 1 = présence d'un risque de crise cardiaque



Instructions d'utilisation

Vous trouverez le fichier de données 'heart.csv' en pièce jointe dans le travail que j'ai soumis. Veuillez le télécharger sur votre appareil.

Recherchez l'onglet 'Fichier' et téléchargez le fichier CSV en appuyant sur le bouton indiqué dans l'image.



Question III.11

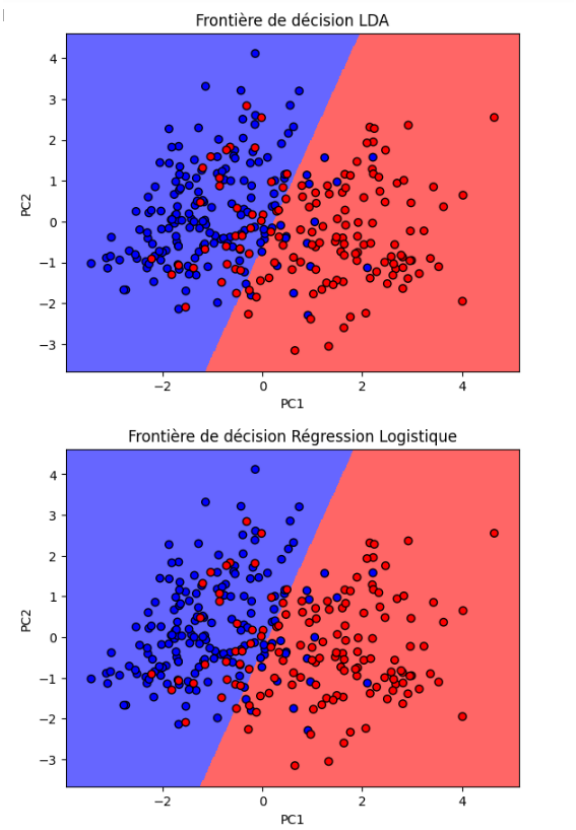
Sur les données d'entraînement, la régression logistique présente une erreur légèrement inférieure à celle de LDA (0,1364 contre 0,1405). Cependant, sur les données de test, la régression logistique affiche une erreur considérablement plus élevée que LDA (0,1475 contre 0,1311).

```
falling loss
beta= [[-0.07952784]
 [-0.8285988 ]
 [ 0.88622068]
 [-0.30907016]
 [-0.19505602]
 [ 0.10466406]
 [ 0.31057468]
 [ 0.42769333]
 [-0.54336128]
 [-0.76385985]
 [ 0.47000958]
 [-0.87528826]
 [-0.61556326]
 [-0.03302207]]
w= [ 0.04669413  0.73897834 -0.86995717  0.29915659  0.10613198 -0.08304313
 -0.22282882 -0.43245102  0.65316938  0.59590601 -0.53391472  0.92134878
  0.62976244] , b= -0.20270390069840702
Error ratio of new model generated by LDA on training data:  0.14049586776859505
Error ratio of new model generated by LDA on testing data:  0.13114754098360656
Error ratio of new model generated by Logistic Regression on training data:  0.13636363636363635
Error ratio of new model generated by Logistic Regression on testing data:  0.14754098360655737
```

Lors de la validation croisée à n blocs, on observe que la moyenne et la variance des taux d'erreur du modèle de régression logistique sont supérieures à celles de LDA également (Moyenne : 0,1749 contre 0,1683 ; Variance : 0,1443 contre 0,140).

```
Mean errors Logistic Regression 0.17491749174917492
Var errors Logistic Regression 0.14432136282935226
Mean errors LDA 0.16831683168316833
Var errors LDA 0.1399862758553083
```

On obtient deux frontières de décision presque similaires



Ce résultat est plutôt surprenant, et je ne peux pas l'expliquer de manière concluante. En théorie, dans les cas où il y a un grand volume d'observations et de variables, les classes ne sont pas bien séparées, et la distribution de X n'est pas approximativement normale **(1)** dans chacune des classes, la régression logistique sera normalement plus efficace.**(2)**

(1): Cette conclusion est **heuristiquement** déduite en observant les matrices de p-value sur les données à deux classes, générées par la méthode `statistic_data`. Si, pour une classe, le nombre de "True" est supérieur au nombre de "False", on déduit que la distribution de X est approximativement normale dans cette classe."

```
def statistic_data(X,y):
  print("Number of observations: ",X.shape[0])
  print("Number of features: ",X.shape[1])
  stat_0, p_value_0 = stats.normaltest(X[y==0])
  stat_1, p_value_1 = stats.normaltest(X[y==1])
  print("Is distribution of X approximately normal in class 0: ",p_value_0>= 0.05)
  print("Is distribution of X approximately normal in class 1: ",p_value_1>= 0.05)
```

```
Is distribution of X approximately normal in class 0:  [False False False False  True False False  True False False  True False
 False]
Is distribution of X approximately normal in class 1:  [ True False False False False False False False False False False False
 False]
```

(2): Source: Page 9, Azad Abdulhafedh, 2022, "Comparison between Common Statistical Modeling Techniques Used in Research, Including: Discriminant Analysis vs Logistic Regression, Ridge Regression vs LASSO, and Decision Tree vs Random Forest," University of Missouri, State of Missouri, USA, https://www.scirp.org/pdf/oalibj_2022021716322673.pdf

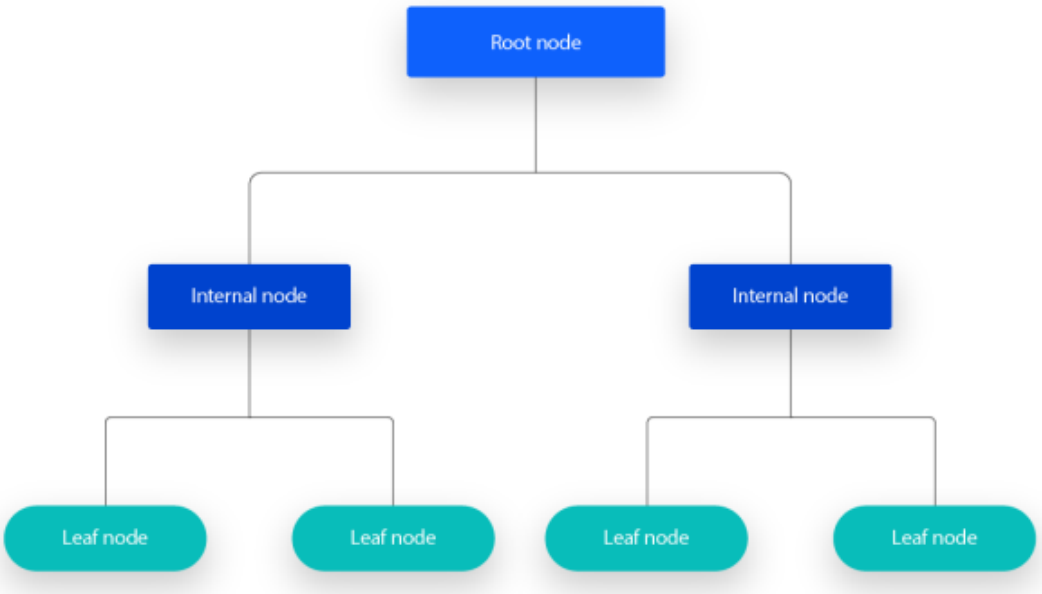
Question III.12

L’approche choisie: Arbres de décisions

*Les **arbres de décision** sont une technique populaire en Machine Learning, utilisée pour la classification et la régression. Voici un aperçu succinct de leur fonctionnement :*

Construction de l'arbre :

- *L'algorithme commence par choisir la meilleure caractéristique (attribut) pour diviser les données d'entraînement en sous-groupes. La "meilleure" caractéristique est celle qui maximise la séparation entre les classes cibles (pour la classification) ou minimise l'erreur de prédiction (pour la régression).*
- *Ce processus de division est répété de manière récursive, créant une structure arborescente où chaque nœud représente une décision basée sur une caractéristique.*



Critères de division :

Les critères couramment utilisés pour mesurer la qualité d'une division comprennent l'entropie, le gain d'information (pour la classification), ou la réduction d'erreur quadratique (pour la régression).

$$\text{Entropy}(S) = -\sum_{c \in C} p(c)\log_2 p(c)$$

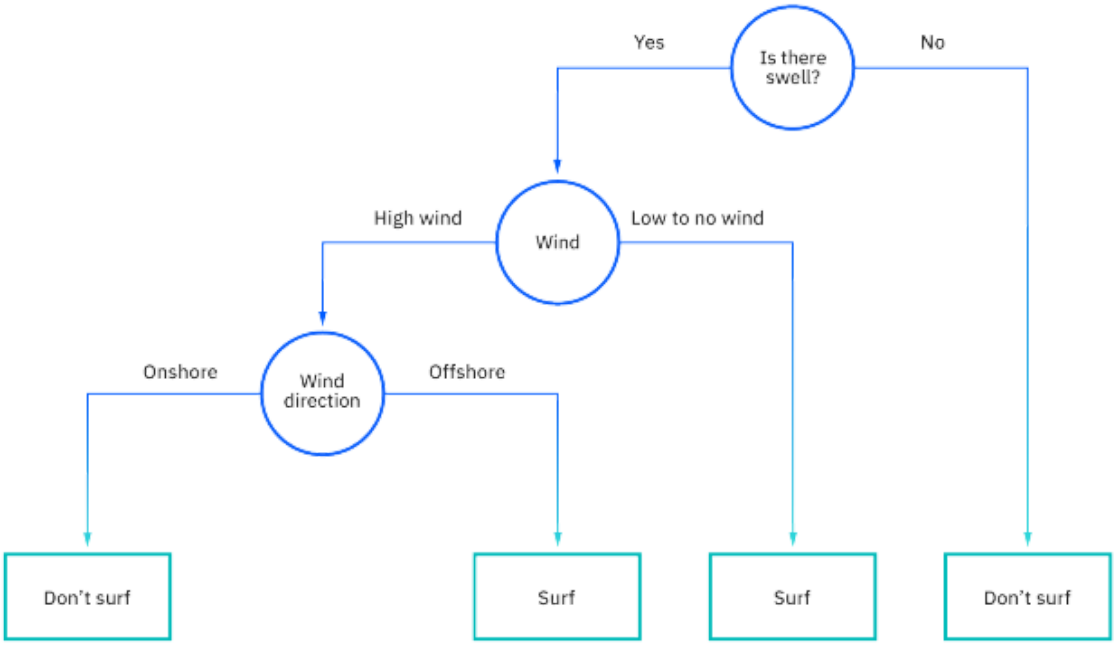
$$\text{Information Gain}(S,a) = \text{Entropy}(S) - \sum_{v \in \text{values}(a)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

Arrêt de la croissance de l'arbre :

La construction de l'arbre se poursuit jusqu'à ce qu'un critère d'arrêt soit atteint, comme une profondeur maximale de l'arbre, un nombre minimum d'échantillons dans une feuille, ou une certaine pureté des feuilles.

Prédiction :

Une fois l'arbre construit, il peut être utilisé pour faire des prédictions sur de nouvelles données en les faisant traverser l'arbre selon les règles apprises pendant l'entraînement. Chaque feuille de l'arbre représente une classe ou une valeur de sortie.



Interprétation :

L'un des avantages des arbres de décision est leur interprétabilité. On peut visualiser un arbre pour comprendre comment les décisions sont prises à chaque étape, ce qui peut être utile pour expliquer les prédictions du modèle.

Les arbres de décision peuvent être utilisés seuls ou en combinaison pour former des structures plus complexes, comme les forêts aléatoires (ensembles d'arbres) pour améliorer la robustesse et la précision du modèle. Ils sont largement utilisés dans divers domaines en raison de leur simplicité, de leur flexibilité et de leur capacité à gérer des données non linéaires et complexes.

Source des images: https://www.ibm.com/fr-fr/topics/decision-trees

Comparer cette 3ème approche avec les 2 autres

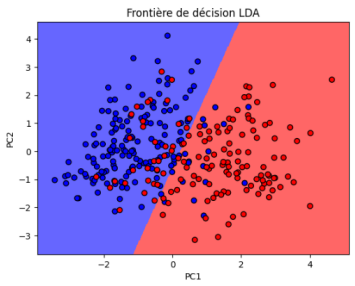
Sur les données de la question 10, pour les données d'entraînement, l'Arbre de décisions affiche un taux d'erreur de 0. Cependant, le taux d'erreur sur les données de test est considérablement plus élevé que celui de LDA et de la régression logistique (0,2131 contre 0,1475 et 0,1311).

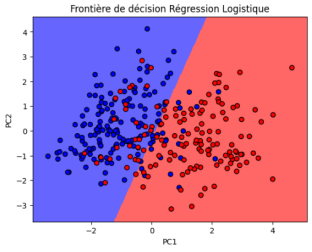
Lors de la validation croisée à n blocs, on observe que la moyenne et la variance des taux d'erreur du modèle de régression logistique sont supérieures à celles de LDA et de la régression logistique également (Moyenne : 0,2145 contre 0,1749 et 0,1683 ; Variance : 0,1685 contre 0,1443 et 0,14).

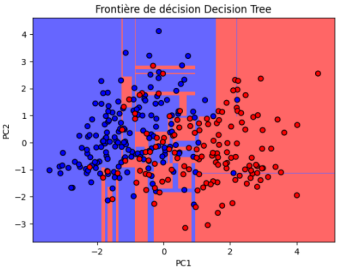
Cela peut être expliqué par la caractéristique de surajustement (Overfitting) de l'algorithme **(3)** : les arbres de décision sont sujets au surajustement, ce qui signifie qu'ils peuvent capturer le bruit et des motifs spécifiques uniques aux données d'entraînement. Lorsqu'un arbre de décision est autorisé à croître trop profondément, il peut créer des règles excessivement complexes qui s'adaptent parfaitement aux données d'entraînement. En effet, l'arbre mémorise essentiellement les données d'entraînement, atteignant une erreur d'entraînement nulle. Cependant, cela peut conduire à une mauvaise généralisation sur de nouvelles données invisibles, car le modèle a essentiellement appris le bruit et les particularités de l'ensemble d'entraînement plutôt que les véritables motifs sous-jacents.

```
Error ratio of new model generated by LDA on training data:  0.14049586776859505
Error ratio of new model generated by LDA on testing data:  0.13114754098360656
Error ratio of new model generated by Logistic Regression on training data:  0.13636363636363635
Error ratio of new model generated by Logistic Regression on testing data:  0.14754098360655737
Error ratio of new model generated by Decision Tree on training data:  0.0
Error ratio of new model generated by Decision Tree on testing data:  0.21311475409836064
Mean errors Logistic Regression 0.17491749174917492
Var errors Logistic Regression 0.14432136282935226
Mean errors LDA 0.16831683168316833
Var errors LDA 0.1399862758553083
Mean errors Decision Tree 0.2145214521452145
Var errors Decision Tree 0.168501998714723
```

On observe que la frontière de décision de l'Arbre de décisions est totalement différente par rapport aux deux autres. Elle n'est pas une ligne, mais plutôt plusieurs lignes. C'est parce que l'Arbre de décisions est un **classificateur non linéaire (2)**. Ces frontières sont parallèles à l'axe est parce qu'un arbre de décision divise les données en fonction d'une valeur de caractéristique, et cette valeur reste constante pour une limite de décision, par exemple, x = 2 ou y = 3, où x et y sont deux caractéristiques différentes. Alors que dans un classificateur linéaire, une limite de décision pourrait être, par exemple : y = mx + c. (3)







(3): Source: Collectivité, 2021, "Random Forest, high test data error but zero training data error", Stack Overflow, https://stackoverflow.com/questions/70020244/random-forest-high-test-data-error-but-zero-training-data-error

(4): Source: Collectivité, 2015, “Are decision tree algorithms linear or nonlinear”, Stack Exchange, https://datascience.stackexchange.com/questions/6787/are-decision-tree-algorithms-linear-or-nonlinear

(5): Source: Azika Amelia, 2019, “Decision tree: Part 1/2”, Medium - Towards Data Science, https://towardsdatascience.com/decision-tree-overview-with-no-maths-66b256281e2b#:~:text=The reason for this is,%3A y%3Dmx%2Bc.

Annexe: Table comparative complète des 3 approches pour 4 datasets

	LDA	Régression Logistique	Arbre de décision
Dataset 1: n=100; p=2; 2 variables informatives; Pas de variables redondantes; Pas de outlier; La distribution de X est approximativement normale dans chaque classe (Heuristiquement)	Taux d'erreur sur données d'entraînement: 0.0375; Taux d'erreur sur données de test: 0.05; Validation croisée à n blocs: - Moyenne: 0.04 - Variance: 0.0384	Taux d'erreur sur données d'entraînement: 0.025; Taux d'erreur sur données de test: 0.05; Validation croisée à n blocs: - Moyenne: 0.04 - Variance: 0.0384	Taux d'erreur sur données d'entraînement: 0.0; Taux d'erreur sur données de test: 0.15; Validation croisée à n blocs: - Moyenne: 0.09 - Variance: 0.0819
Dataset 2: n=100; p=2; 2 variables informatives; Pas de variables redondantes; Un outlier; La distribution de X est approximativement normale dans chaque classe (Heuristiquement)	Taux d'erreur sur données d'entraînement: 0.05; Taux d'erreur sur données de test: 0.05; Validation croisée à n blocs: - Moyenne: 0.04 - Variance: 0.0384	Taux d'erreur sur données d'entraînement: 0.0375; Taux d'erreur sur données de test: 0.05; Validation croisée à n blocs: - Moyenne: 0.04 - Variance: 0.0384	Taux d'erreur sur données d'entraînement: 0.0; Taux d'erreur sur données de test: 0.15; Validation croisée à n blocs: - Moyenne: 0.09 - Variance: 0.0819
Dataset 3: n=100; p=7; 2 variables informatives; 2 variables redondantes; Pas d'outlier; La distribution de X est approximativement normale dans chaque classe (Heuristiquement)	Taux d'erreur sur données d'entraînement: 0.0375; Taux d'erreur sur données de test: 0.15; Validation croisée à n blocs: - Moyenne: 0.1277 - Variance: 0.1114	Taux d'erreur sur données d'entraînement: 0.0125; Taux d'erreur sur données de test: 0.1; Validation croisée à n blocs: - Moyenne: 0.07 - Variance: 0.0651	Taux d'erreur sur données d'entraînement: 0.0; Taux d'erreur sur données de test: 0.05; Validation croisée à n blocs: - Moyenne: 0.08 - Variance: 0.0736
Dataset 4: n=303; p=13; ? variables informatives; ? variables redondantes; Pas d'outlier; La distribution de X n'est pas normale dans	Taux d'erreur sur données d'entraînement: 0.1405; Taux d'erreur sur données de test: 0.1311; Validation croisée à	Taux d'erreur sur données d'entraînement: 0.1364; Taux d'erreur sur données de test: 0.1475; Validation croisée à	Taux d'erreur sur données d'entraînement: 0.0; Taux d'erreur sur données de test: 0.2131; Validation croisée à n blocs: -

chaque classe (Heuristiquement)	n blocs: - Moyenne: 0.1683 - Variance: 0.14	n blocs: - Moyenne: 0.1749 - Variance: 0.1443	Moyenne: 0.2145 - Variance: 0.1685