

**Masters SEP & Calcul Scientifique**

**2022-2023**

**Cours d’apprentissage automatique**

**Projet final**

**L’analyse discriminante quadratique (QDA)**

**Réalisé par :**

Laitissia ALI (2)

Ismael Djoulde DIALLO (1 & 4.3)

Alizée SCHOLLORN-LEOPOLD (3)

Yacine ZIDI (4)

Table des matières

[1. Présentation de l’analyse discriminante quadratique (the QDA) 3](#_Toc127741087)

[2. Mécanismes mathématiques mis à l’œuvre 4](#_Toc127741088)

[2.1 Contexte : idées générales 4](#_Toc127741089)

[2.2 Difficulté : 4](#_Toc127741090)

[2.3 Idée de la mise en œuvre : 4](#_Toc127741091)

[2.4 Aspects Mathématiques : 5](#_Toc127741092)

[2.4.1 Hypothèses sur les données avant application de la méthode : 5](#_Toc127741093)

[2.4.2 Pour séparer les classes : 5](#_Toc127741094)

[2.4.3 Ajustement des paramètres de la courbe : 5](#_Toc127741095)

[2.4.4 Evaluation de la performance du modèle : 6](#_Toc127741096)

[3. Implémentation de la méthode QDA sous Python 7](#_Toc127741097)

[4. Comparaison du QDA avec d'autres algorithmes de classification supervisées : 10](#_Toc127741098)

[4.1 L’analyse discriminante linéaire 10](#_Toc127741099)

[4.2 La régression logistique 10](#_Toc127741100)

[4.3 QDA vs Regularized discriminant analysis (RDA) 12](#_Toc127741101)

[4.3.1 Vérification des hypothèses 12](#_Toc127741102)

[4.3.2 Comparaison des résultats de la QDA et la RDA 15](#_Toc127741103)

[Références 17](#_Toc127741104)

# Présentation de l’analyse discriminante quadratique (the QDA)

L'analyse discriminante est un ensemble de méthodes de classification supervisée qui vise à déterminer les caractéristiques les plus discriminantes pour séparer des groupes d'individus préalablement définis. Soulignons toutefois qu’il existe des versions non supervisées de l’analyse discriminante comme Un-LDA[[1]](#footnote-1) que nous n’aborderons pas ici. Dans ce travail, nous nous intéressons à l’analyse discriminante quadratique.

L’analyse discriminante quadratique est une technique d’analyse discriminante non linéaire et paramétrique permettant de prédire l’appartenance d’une observation à une classe parmi plusieurs classes différentes définies a priori à partir de ses caractéristiques mesurées à l’aide de variables explicatives. L'analyse discriminante quadratique est dite paramétrique parce qu’elle suppose que les données d'apprentissage proviennent de distributions gaussiennes multivariées distinctes. Elle est dite non linéaire, contrairement à l’analyse discriminante linéaire LDA parce qu’elle s’applique à des données ne pouvant être séparées que par une courbe quadratique pour construire la frontière qui sépare les classes.

Considérons un jeu de données constitué de n observations qui sont réparties dans K classes de tailles nk, de centre de gravité μk et de matrice variance-covariance Γk.Nous souhaitons prédire une variable Y ayant K modalités caractérisant les K classes à partir de de J variables prédictives X pouvant être numériques ou catégorielles[[2]](#footnote-2). Le problème revient donc à déterminer une fonction d’affectation qui permettrait de classer toute observation de paramètres X1, .., Xj dans l’une des K classes prédéfinies.

Pratiquement, l’ADQ consiste à modéliser les probabilités de chacune des K classes, ensuite pour une observation donnée, l'algorithme calcule la probabilité qu'elle appartienne à chaque classe en utilisant les distributions de probabilité de chaque classe. Ainsi, la classe avec la probabilité la plus élevée est choisie comme étiquette de classe pour cette observation.

L’analyse discriminante quadratique à un large champ d’application. On a recours à la QDA pour des problèmes de classification sur des données gaussiennes reparties en des classes distinctes de variances et covariances différentes. Elle est utilisée dans la détection de fraudes, le scoring, la reconnaissance d’objets, dans les problèmes de classification médicale, identification des origines génétiques d’un individu, comparaison des traces d’ADN trouvées sur une scène de crime à une base de données d’ADN.

Par ailleurs, on ne pourrait parler de l’ADQ sans parler de l’analyse discriminante régularisée RDA. La RDA permet de traiter problèmes de surajustement de l’ADQ qui se produisent lorsque le nombre de variables explicatives augmente. En plus des hypothèses de l’ADQ suivants :

* Les observations sont indépendantes les unes des autres ;
* Chacune des classes suit une distribution gaussienne multivariée ;
* Les groupes ont des covariances différentes ;
* Les variables explicatives sont linéairement indépendantes les unes des autres ;

Une hypothèse supplémentaire liée à la régularisation est introduite dans l’analyse discriminante régularisée. De fait, cette hypothèse traduit l’usage d’un coefficient de pénalisation dans la fonction d’optimisation utilisée pour estimer les paramètres du modèle.

Dans les parties 2, 3 et 4, nous aborderons plus en détail les concepts présentés ici.

# Mécanismes mathématiques mis à l’œuvre

## Contexte : idées générales

On se donne un jeu de données, puis on cherche à classer les observations que l’on dispose en créant différents groupes.

Pour les explications qui suivent, en guise d’illustration, on se basera parfois sur le jeu de données « **iris »** ayant pour variables : les mesures de longueur et de largeur de pétales et de sépales pour trois espèces de fleurs différentes (Iris setosa, Iris versicolor et Iris virginica), donnant au total 5 variables.

On peut utiliser la méthode de la QDA pour classer ces fleurs en différentes catégories selon les espèces. Les différentes catégories seraient les différentes modalités de la variable « **espèce »** du jeu de donnée « **iris »**.

La QDA est une méthode permettant de déterminer à quelle catégorie appartient un individu (une observation) en fonction de ses caractéristiques.

L’idée est de savoir, à partir de la longueur des pétales et de sépales, la largeur des pétales et des sépales, donner l’espèce à laquelle la fleur appartient.

## Difficulté :

Après avoir prédéfinis les catégories, comment trouver une courbe qui sépare les différents groupes en fonction des caractéristiques des observations ? Comment juger de la qualité (la performance) de cette courbe ?

## Idée de la mise en œuvre :

On commence par une étape de prédiction qui consiste à considérer les autres variables de la base comme des variables prédictives auxquelles on a posé une étiquette de classes connues. La variable cible est celle qui donne les classes que l’on souhaite séparer.

On divise ensuite le jeu de données en un ensemble d’entraînement et un ensemble de test. L’ensemble d’entraînement est utilisé pour ajuster les paramètres de la courbe de séparation que l’on veut créer. L’ensemble de test est utilisé pour évaluer les performances du modèle i.e la capacité à déterminer sans erreur l’espèce de la fleur (:= la classe de l’observation) en fonction des caractéristiques que l’on a fourni en entrée.

L’ajustement des paramètres maximise la précision de la classification sur les données d’apprentissage.

La courbe obtenue pourrait être utilisée pour prédire la catégorie d’une nouvelle observation pas encore enregistrée dans le jeu de donnée. Avec le jeu de données « iris », la courbe peut permettre de prédire l’espèce d’une nouvelle fleur (pas encore enregistrée dans le jeu de données) en fonction de ses mesures de pétales et de sépales.

## Aspects Mathématiques :

La QDA pour la classification n’est pas un algorithme géométrique. L’essentiel de la méthodologie ne repose pas sur les distances entre observations.

Cette méthode utilise des modèles de probabilité pour estimer les probabilités d’appartenance à chaque classe. Elle repose sur des hypothèses probabilistes sur la distribution des données d’apprentissage dans chaque classe.

### Hypothèses sur les données avant application de la méthode :

La méthode QDA suppose que :

* Les données dans chaque classe suivent une distribution normale multivariée. i.e pour chaque classe, les valeurs de chaque variable sont normalement distribuées, et les différentes variables sont corrélées entre elles.
* Les différentes classes ont des matrices de covariance différentes i.e les données dans chaque classe peuvent avoir des formes différentes, et les relations entre les différentes variables peuvent varier d’une classe à l’autre. La matrice de covariance est utilisée pour décrire la relation entre les différentes variables.

### Pour séparer les classes :

Le théorème de Bayes est utilisé pour estimer la probabilité d’appartenance d’une observation à un groupe donné. Ce théorème permet de combiner les probabilités a priori et les probabilités conditionnelles pour obtenir la probabilité à posteriori.

La décomposition en valeurs propres : est utilisée pour calculer les axes principaux de la distribution des données. Ces axes principaux peuvent être utilisés pour réduire la dimension des données et pour visualiser les différences entre les groupes de données.

### Ajustement des paramètres de la courbe :

La QDA cherche une courbe séparant les classes en maximisant la probabilité d’attribution d’un point à sa classe réelle, en fonction des données d’entraînement. Pour trouver cette courbe séparant les classes, la QDA calcule une fonction discriminante quadratique pour chaque classe ; c’est une équation quadratique en les variables prédictives. Ensuite, la méthode trouve la frontière de décision en comparant les fonctions discriminantes pour chaque paire de classes et en trouvant les points où les fonctions sont égales.

Pour modéliser les classes, elle utilise des fonctions de densité de probabilité. Pour chaque classe, on calcule la densité de probabilité en utilisant une distribution gaussienne multivariée.

L’ajustement du modèle se fait grâce à la méthode des moindres carrés. Elle consiste à minimiser la somme des carrés des résidus entre les données observées et les valeurs prédites par le modèle.

Les paramètres du modèle sont ajustés en maximisant la fonction de vraisemblance. Cette fonction mesure la probabilité de voir les données observées étant donné les paramètres du modèle. Les paramètres sont ajustés pour minimiser la distance entre les distributions de probabilité de chaque groupe et maximiser la distance entre les centres de gravité des distributions de probabilité de chaque groupe. Cela permet de déterminer les limites de décision qui séparent les groupes, en utilisant des courbes quadratiques.

Dans le cadre de la méthode QDA pour la classification, on cherche à maximiser la vraisemblance de la solution. Pour cela, on utilise l’algorithme EM (Expectation-Maximization) qui est une procédure itérative de maximisation de la vraisemblance. Cet algorithme permet d’optimiser les paramètres de la distribution de probabilité des données pour chaque classe, en utilisant les données d’entraînement disponibles.

Une fois les paramètres estimés, la solution est évaluée en utilisant la performance de classification sur un ensemble de données de test. L’objectif est de trouver la solution qui minimise l’erreur de classification sur cet ensemble de données.

La matrice de covariance peut être utilisée pour estimer les paramètres du modèle et pour calculer la distance entre les différentes observations.

### Evaluation de la performance du modèle :

La performance du modèle est mesuré en calculant le taux de point mal classés (i.e erreur de classification). On va donc utiliser la matrice de confusion afin d’évaluer nos classifications. Cette matrice carrée affiche le nombre d’observations classées dans chaque catégorie en fonction de la catégorie réelle à laquelle elles appartiennent.

À partir de cette matrice, on peut calculer plusieurs métriques de performance, telles que la précision, le rappel ou le F1-score.

La **précision** est donnée par le nombre d’observations correctement classées dans une catégorie donnée, divisé par le nombre total d’observations classées dans cette catégorie.

Le **rappel** est le nombre d’observations correctement classées dans une catégorie donnée, divisé par le nombre total d’observations réelles dans cette catégorie.

Le **F1-score** est la moyenne harmonique de la précision et du rappel.

En maximisant la précision et le rappel pour chaque catégorie, cela revient à minimiser le taux d’erreurs de classification.

# Implémentation de la méthode QDA sous Python

Sous Python, la méthode est déjà implantée, elle se trouve dans la librairie sklearn.discriminant\_analysis. Cela est très pratique puisqu’on peut l’utiliser directement.

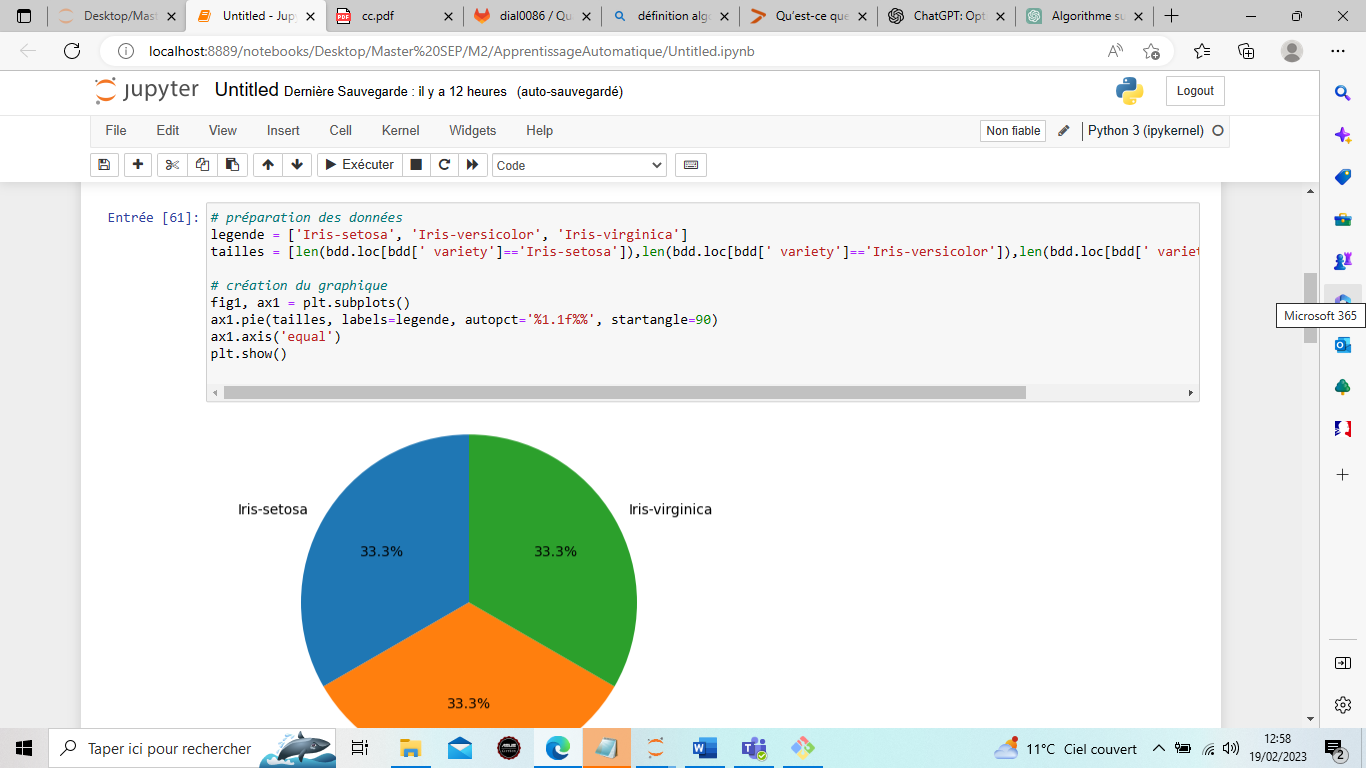
Etant donné que c’est un algorithme de classification supervisé[[3]](#footnote-3), il est nécessaire d’avoir un nombre similaire de données dans chaque catégories. Ici, sur le graphique circulaire, on obtient 33.3% de données pour chacune des catégories d’iris, il n’y a donc pas besoin de faire un rééquilibrage des classes. Si ce problème avait eu lieu, on aurait eu des résultats biaisés en faveur de la classe majoritaire. De plus, la normalisation des données n’a pas été effectuée puisqu’elles sont déjà toutes dans la même unité. Après ces vérifications, il est impératif de diviser notre base de données en deux parties, d’entraînement et de test comme nous l’avons vu précédement. En effet, l’algorithme va d’abord apprendre les données sur la base d’entraînement pour ensuite les tester sur la base de test. Si on ne fais pas ce découpage en deux parties distinctes, l’algorithme va apprendre sur toutes nos données puis tester sur 20% de ces mêmes données. On aura alors d’excellents résultats sur notre phase de test. Or, l’algorithme à déjà appris sur ces données-ci, c’est donc normal qu’il connaisse la prédiction exacte mais sur de nouvelles données, ce ne sera pas le cas, et on aura de très mauvais résultats.

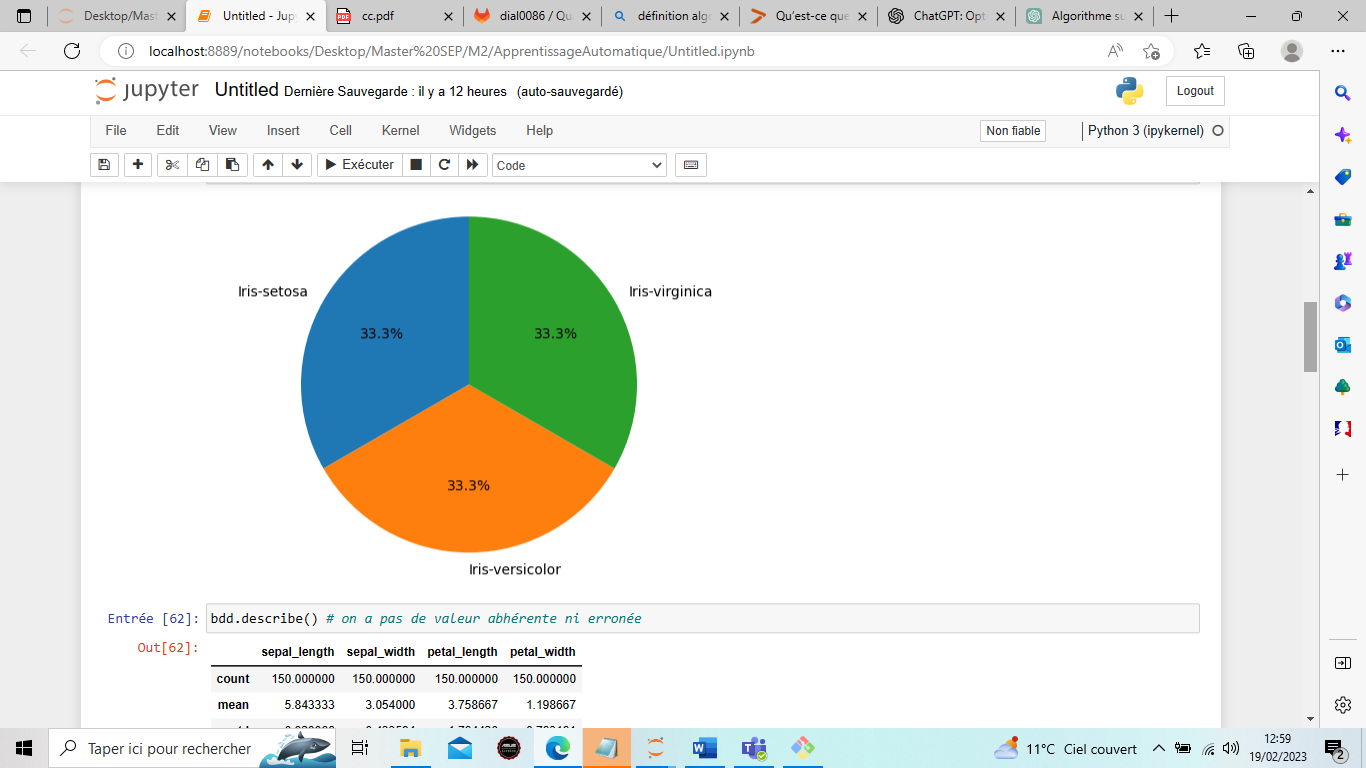
Une fois qu’on a prédit les résultats sur les 20% de notre base de test, ils sont ensuite comparés aux résultats attendus afin d’avoir un pourcentage de bonne prédiction de notre modèle. Cela nous permettra donc de savoir si ce modèle est fiable ou non.

Lors de notre premier test, nous avons obtenu une précision de 1 pour notre modèle, ce qui est excellent. Cependant, afin d’être sûr de qualité de notre modèle, nous l’avons réitéré 1000 fois, en changeant de découpage systématiquement, pour ne pas utiliser les mêmes données pour la phase d’apprentissage. La moyenne de ces 1000 précisions est de 0.97 ce qui reste toujours excellent. On peut donc en conclure que ce modèle QDA prédit très bien nos données sur les iris.

Une image contenant texte

Description générée automatiquement





Une image contenant texte

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

Une image contenant texte

Description générée automatiquement

# Comparaison du QDA avec d'autres algorithmes de classification supervisées :

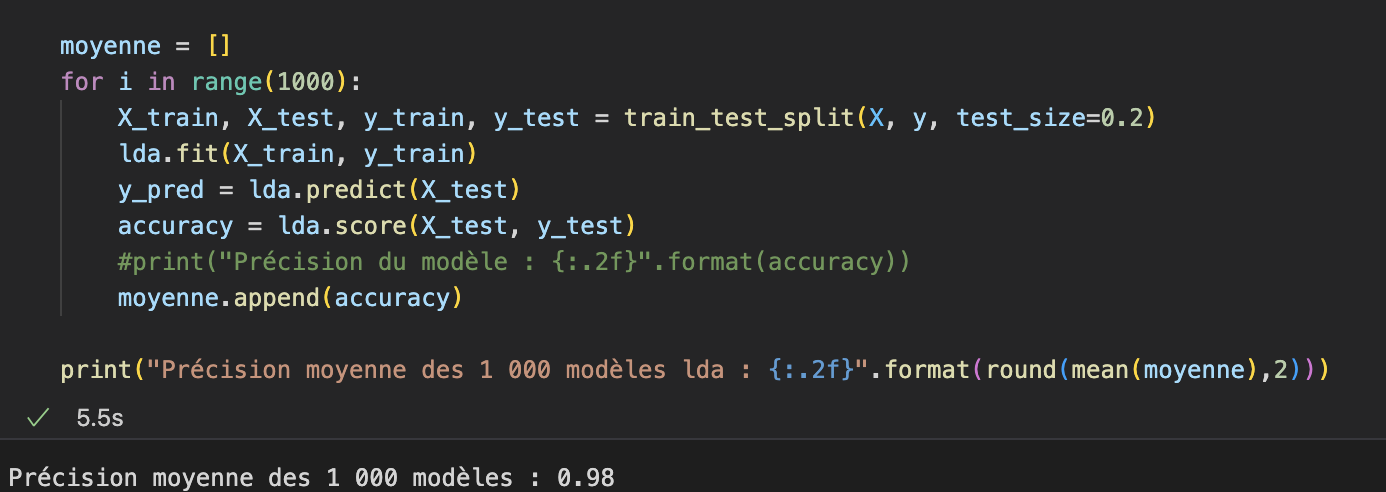
 Dans cette partie, nous allons comparer le modèle de QDA ci-dessus avec trois autres modèles de classifications supervisées : la régression logistique, l'analyse discriminante linéaire et l’analyse discriminante régularisée. Pour comparer ces modèles, nous allons nous concentrer sur le critère de précision. Autrement dit, est-ce que ces autres modèles sont aussi performant que le QDA en termes de prédiction.

## L’analyse discriminante linéaire

L'analyse discriminante linéaire diffère de la QDA par la technique de séparations des groupes. Cette frontière est modélisée par une fonction linéaire, autrement dit une droite qui va séparer au mieux en minimisant l'erreur de classification les différents groupes. Elle s'appuie également sur des données gaussiennes.

La fonction de "classification" utilisée en QDA est dite quadratique dans la mesure ou elle peut être conique et donc non-linéaire. Ainsi, la QDA modélise les relations plus complexes entre les caractéristiques et les classes.

Nous avons donc implémenter cette méthode, toujours sur la base iris, ainsi qu’avec une base d’entraînement sur 80% de nos données et une base de test sur les 20% restantes. Ici, suite à 1 000 implémentations de la méthode, nous obtenons une précision de 0.98.

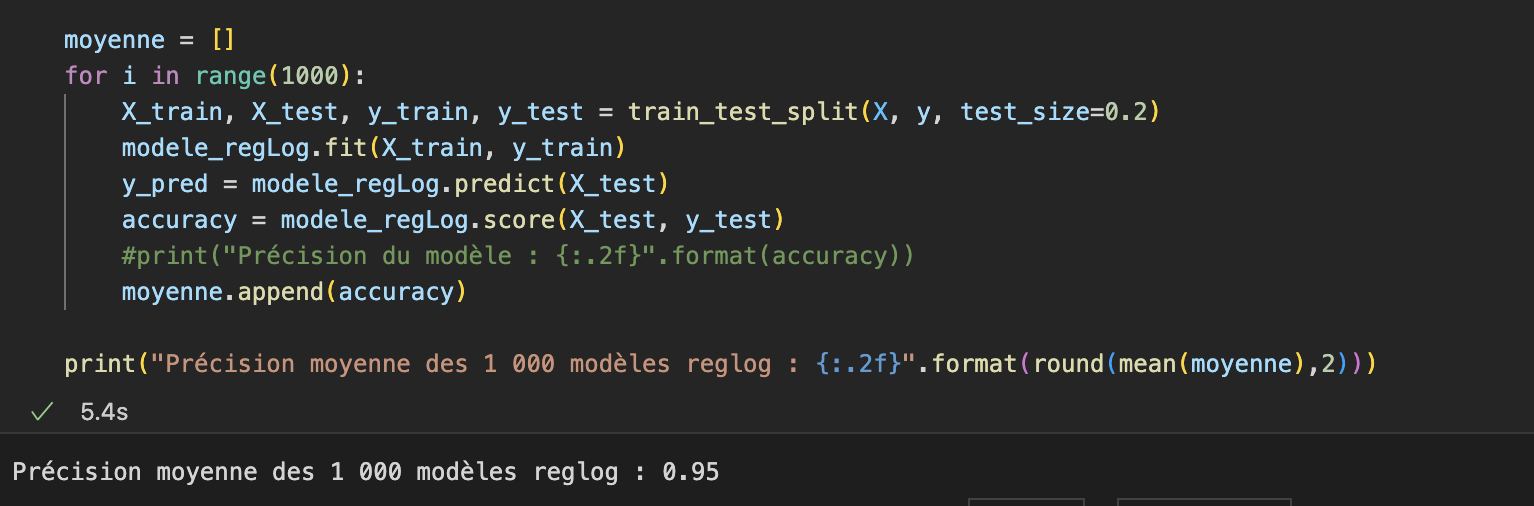
****

## La régression logistique

En ce qui concerne la régression logistique, cet algorithme dit de classification cherche à séparer les groupes avec une fonction logistique et ainsi trouver la probabilité qu'un individu appartienne à une classe ou à une autre.

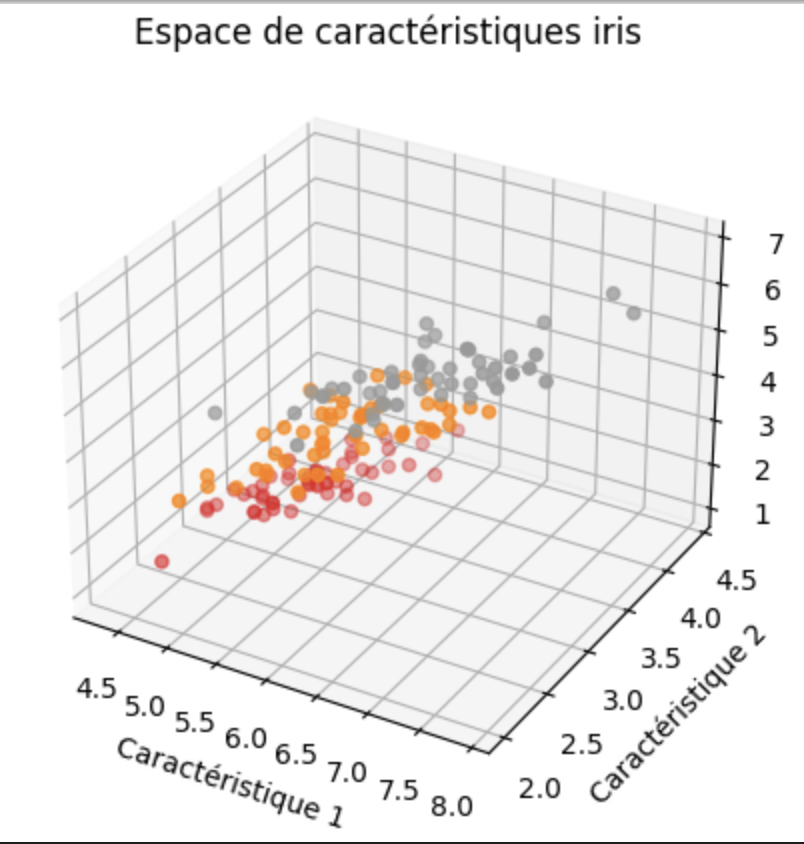
La régression logistique nous permet d’obtenir une précision de 0.95.

On peut donc dire que le LDA dans notre cas a une meilleure précision sur la prédiction de nos données et donc la classification des groupes.

****

Même si la précision des modèles sont très élevées, l'écart entre la LDA,QDA et la régression logistique peut s'expliquer par le fait que la régression logistique ne prend pas en compte l'hypothèse de normalité de la distribution des données par rapport à la QDA et LDA .

Cela peut s'expliquer aussi par le fait que les classes sont déjà bien séparées. En effet, lorsque l'on représente les caractéristiques sur un graphique on peut dessiner 3 groupes distincts.



Ici on a représenté nos individus par 3 caractéristiques ( sepalwidth, sepallenght, petalwidth) et on peut d'ores et déjà séparer en 3 groupes. Lorsque les classes sont déjà bien séparées, les modèles LDA et QDA sont bien plus performants que la régression logistiques.

## QDA vs Regularized discriminant analysis (RDA)

Dans la partie 1 de ce travail, nous avons montré que la RDA permettait d’éviter le surajustement de la QDA. Ici, nous montrons comment cela fonctionne sur des données réelles. Pour ce faire, nous utilisons le jeu de données Pima Indians Diabetes téléchargée depuis <https://data.world/data-society/pima-indians-diabetes-database>. Il s’agit de données de diagnostiques de femmes indiennes permettant de savoir si elles sont diabétiques ou non, on a donc a priori deux classes.

Après nettoyage des données, nous retenons 336 patientes pour les variables grossesses, taux de glucose dans le sang, tension artérielle, épaisseur de la peau, insuline, indice de masse corporelle BMI, la fonction de prédisposions au diabète[[4]](#footnote-4), l’âge et le résultat du diagnostic.

Une image contenant table

Description générée automatiquement

Figure : Extrait des 10 premières observations du jeu de données.

Nous allons réaliser dans un premier temps, la QDA sur le jeu de données et puis une RDA pour enfin comparer les résultats des méthodes.

Pour effectuer la QDA, nous allons vérifier au préalable que les données sont conformes aux hypothèses de la QDA présentées au point [[1]](#Presentation).

### Vérification des hypothèses

►Les réponses des femmes interrogées sont indépendantes les unes des autres. L'observation d'une variable pour un individu donné ne dépend pas de l'observation de cette même variable pour un autre individu. Que l’une soit diabétique n’influence pas qu’une autre le sera ou ne le sera pas ; Aussi, ce n’est pas parce que l’une a un indice de masse corporelle élevée que l’autre sera dans le même cas.

► Pour vérifier l’hypothèse de normalité des distributions des variables, nous utilisons le diagramme quantile-quantile (QQ-plot).

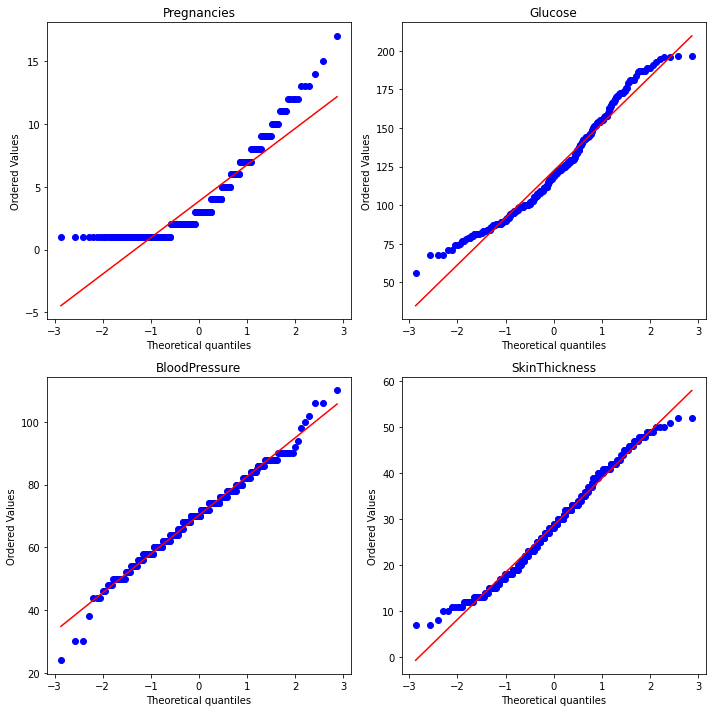


Figure : QQ-plot des 4 premières variables du jeu de données.

Le QQ-plot représente les quantiles d’une variable donnée en abscisse (sur l'axe des X) et les quantiles de la distribution la normale en ordonnée (sur l'axe des Y). Si les données suivent exactement la distribution de référence, les points dans le graphique QQ-plot se situeront sur une droite diagonale, appelée droite de référence ou droite d'égalité (en rouge ici). Si les points s'écartent de cette droite, cela indique que la distribution des données n'est pas tout à fait identique à la distribution de référence. Pour ces 4 variables, c’est seulement la variable qui ne suit pas une loi normale. Les autres variables sont présentées en annexe.

► Pour vérifier l’hypothèse de l’inégalité des variances de la classe des diabétiques et celle des non-diabétiques, on peut utiliser des boites à moustaches ou des ellipses des nuages des deux classes.

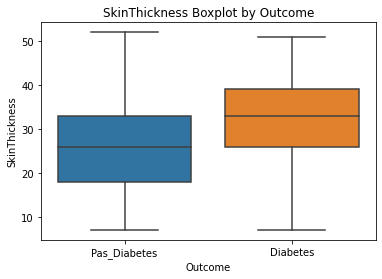
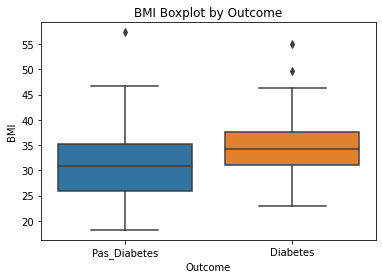


Figure : Boite à moustache des variable SkinThickness et BMI vs la variable Outcome.

On constate que les 2 boites sont différentes et donc que les variances des deux classes pour cette variable sont différentes. Voir les autres boites à moustache dans le notebook ‘pima\_QDA&RDA’.

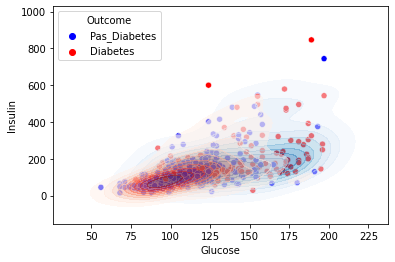


Figure : Nuage de points des variables Insulin et Glucose.

On voit dans ce graphique que les points rouges sont plus dispersés que les points bleus. Les variances des deux classes sont aussi différentes pour ces variables.

► Pour vérifier l’hypothèse d’indépendance des variables, on utilise une matrice de corrélation.

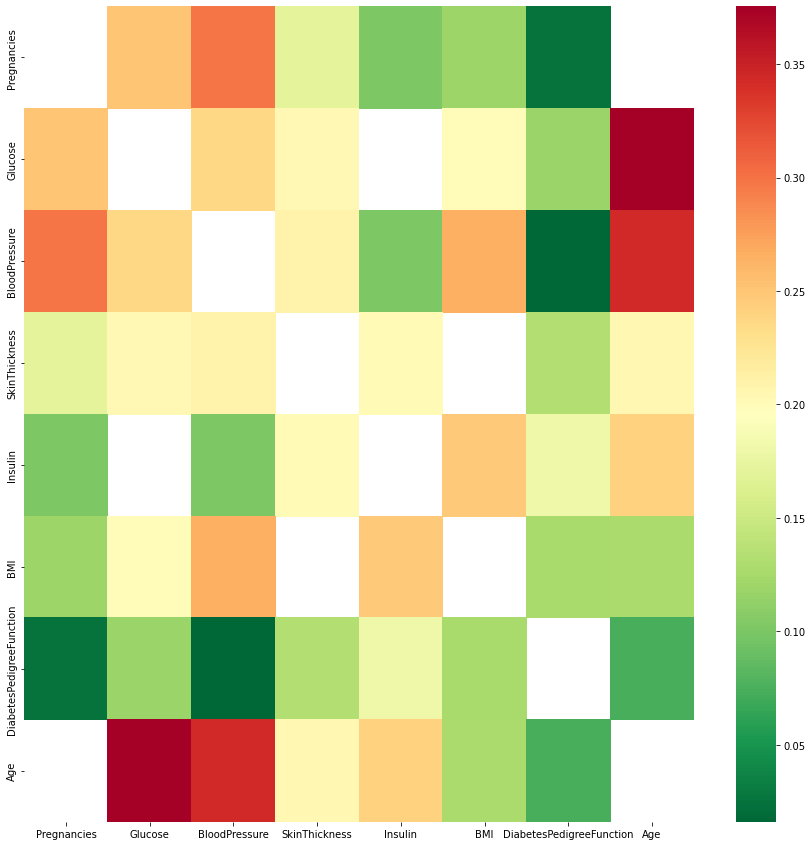


Figure : Matrice de corrélation entre les variables.

Les variables Glucose et BloodPressure sont fortement corrélées avec l’âge (couleurs rouge et rouge foncé) mais le coefficient de corrélation reste tout de même inférieur à 0.5.

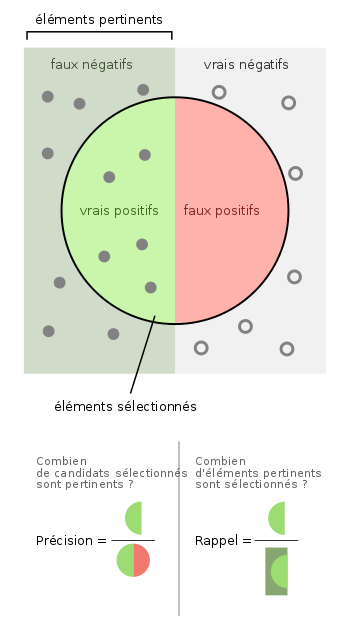
Nous avons vérifié toutes les hypothèses, on retient finalement comme variables explicatives du diabète pour l’analyse discriminante quadratique les variables BMI, Glucose, BloodPressure et SkinThickness parce qu’elles respectent toutes les hypothèses.

### Comparaison des résultats de la QDA et la RDA

Pour effectuer la RDA, il a fallu chercher un coefficient de régularisation qui permet d’éviter le surajustement de la QDA. La technique de validation croisée KFold a donc été utilisé.

|  |  |  |  |
| --- | --- | --- | --- |
| **Modèle** | **Rappel(Recall)** | **Précision** | **Accuracy** |
| **QDA** | **Diabétique : 67%**  **Non Diabétique : 82%** | **Diabétique : 67%**  **Non Diabétique : 82%** | **76%** |
| **RDA** | **Diabétique : 62%**  **Non Diabétique : 82%** | **Diabétique : 65%**  **Non Diabétique : 80%** | **75%** |

Tableau comparatif de la QDA et la RDA.

[[5]](#footnote-5)

Vrais positifs : femmes réellement diabétiques.

Faux positifs : femmes qui ne sont réellement pas diabétiques.

Dans le jeu de données les deux classes ne sont pas équilibrées, nous avons 111 diabétiques et 225 non-diabétiques, on utilise le rappel pour sélectionner le meilleur modèle, l’accuracy serait la métrique à utiliser dans le cas où les données seraient équilibrées.

D’après les résultats dans le tableau comparatif [ci-dessus](#QDA_RDA), la QDA est plus performante que la RDA et donc que la QDA a une meilleure capacité de discrimination des diabétiques et non diabétiques que la RDA. De fait, les bonnes performances de la QDA pourraient être due à un surajustement ou à un sous ajustement, la figure [ci-dessous](#QDA_vs_RDA) nous permettra de mieux appréhender la sensibilité de la QDA aux nouvelles données.

****

Figure : QDA vs RDA

Précisons que le score utilisé ici est la ‘balanced accuracy’ parce que les données sont déséquilibrées.

Si la taille de l'ensemble d'apprentissage est petite, les performances des classifieurs sur cet ensemble d'apprentissage sont élevées. En revanche, à mesure que la taille de l'ensemble d'apprentissage augmente, les performances du modèle sur l'ensemble d'apprentissage diminuent. On constate une variance élevée (surajustement) dans un modèle de prédiction s'il fonctionne bien sur l'ensemble d'apprentissage mais mal sur de nouvelles données.

Pour la QDA, nous constatons que la courbe de validation présente un score inférieur à celui de la courbe d’entrainement cela signifie que le modèle devient sur adapté aux données, de plus les deux courbes ne convergent pas vers le même score. Cependant, pour la RDA on constate que les deux courbes convergent vers le même score, il y a équilibre entre variance et biais.

On conclut donc que les bonnes performances de la QDA étaient dues en réalité à un surapprentissage. Ainsi, la RDA se révèle être la meilleure méthode pour discriminer les diabétiques des non-diabétiques pour [ces données](#donnees_clean).

# Références

Trevor Hastie, Robert Tibshirani, Jerome Friedman. 2009,The Elements of Statistical Learning. 2nd  
edition. Springer. Chap 4.

Jerome H. Friedman. 1989, Regularized Discriminant Analysis. Journal of the American Statistical Association.

Christopher Bishop. 2006, Pattern Recognition and Machine Learning. Springer. Chap 4.

M. Kuhn and K. Johnson. 2013, Applied Predictive Modeling. Springer. Chap 13.

Kevin P. Murphy. 2012, Machine Learning: A Probabilistic Perspective. The MIT press. Chap 4.

1. F. Wang, Q. Wang, F. Nie, Z. Li, W. Yu and R. Wang, "Unsupervised Linear Discriminant Analysis for Jointly Clustering and Subspace Learning," in IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 3, pp. 1276-1290, 1 March 2021, doi: 10.1109/TKDE.2019.2939524. [↑](#footnote-ref-1)
2. Il faudra plutôt utiliser un traitement particulier, avec le One Hot Encoding, créer autant de variables binaires qu’il y’a de modalités pour chaque variable catégorielle. [↑](#footnote-ref-2)
3. Un algorithme supervisé est un type d'algorithme qui apprend à partir d'un ensemble de données d'entraînement étiquetées. [↑](#footnote-ref-3)
4. La fonction de pédiatrie du diabète évalue la probabilité d’avoir le diabète en fonction des antécédents familiaux du diabète. [↑](#footnote-ref-4)
5. <https://leandeep.com/evaluer-ses-mod%C3%A8les-de-classification/> [Consulté le 2/19/2023] [↑](#footnote-ref-5)