

**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 & ADQ vs RDA)

Alizée SCHOLLORN-LEOPOLD (3)

Yacine ZIDI (4)

# 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.

## 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[[3]](#footnote-3), 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).

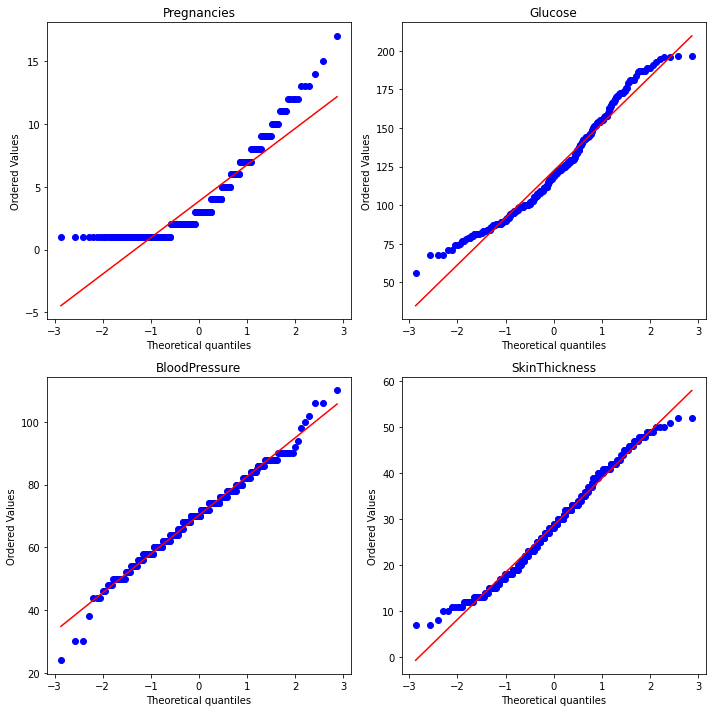


Figure : QQ-plot des 4 premières variables du jeu de données.

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.

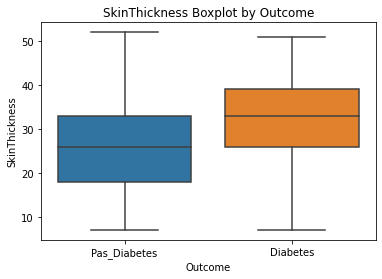
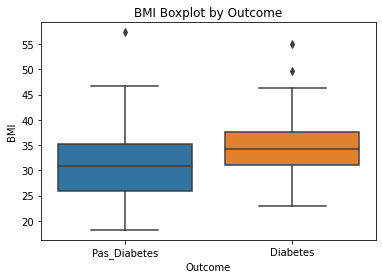


Figure : Boite à moustache des variable SkinThickness et BMI vs la variable Outcome.

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’.

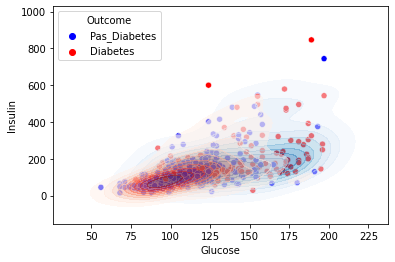


Figure : Nuage de points des variables Insulin et Glucose.

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.

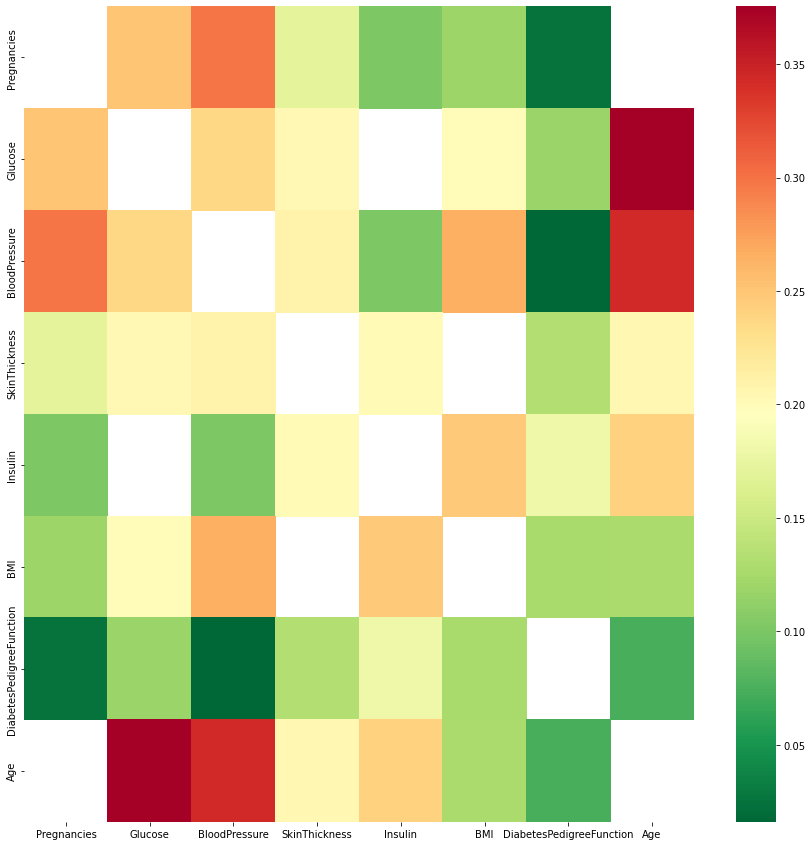


Figure : Matrice de corrélation entre les variables.

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.

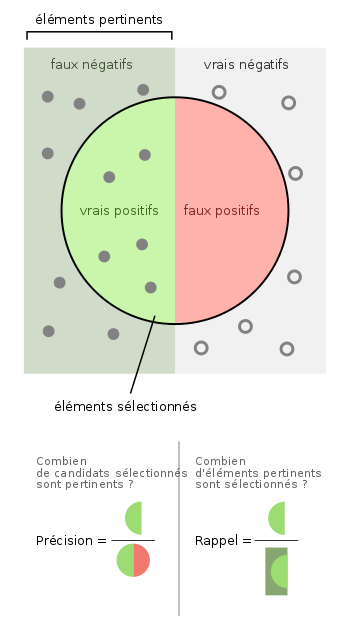
Nous avons vérifié toutes les hypothèses, on retient finalement comme variables explicatives du diabète pour l’analyse discriminante quadratique les variables BMI, Glucose, BloodPressure et SkinThickness parce qu’elles respectent toutes les hypothèses.

### 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.

[[4]](#footnote-4)

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.

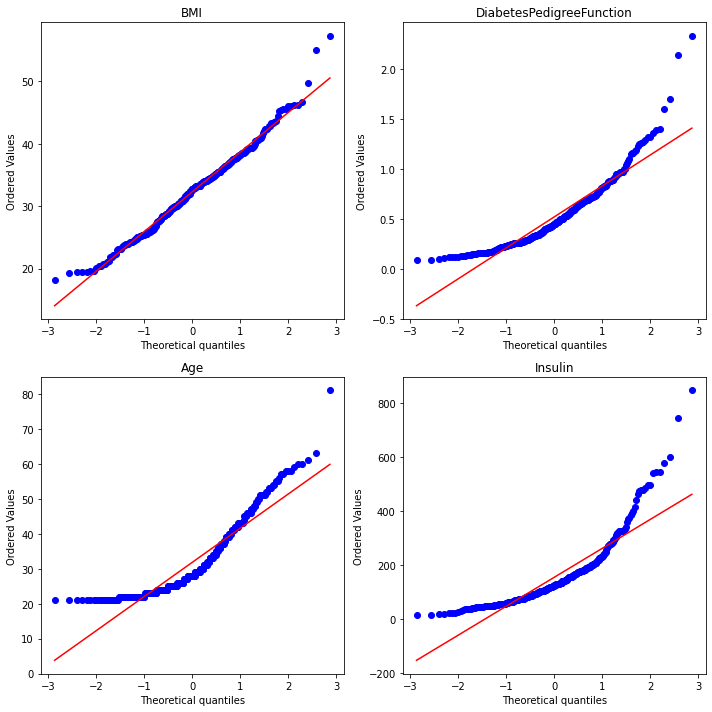
Si la taille de l'ensemble d'apprentissage est petite, les performances des classifieurs sur cet ensemble d'apprentissage sont élevées. En revanche, à mesure que la taille de l'ensemble d'apprentissage augmente, les performances du modèle sur l'ensemble d'apprentissage diminuent. On constate une variance 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).

# Annexe

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.



# 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. La fonction de pédiatrie du diabète évalue la probabilité d’avoir le diabète en fonction des antécédents familiaux du diabète. [↑](#footnote-ref-3)
4. <https://leandeep.com/evaluer-ses-mod%C3%A8les-de-classification/> [Consulté le 2/19/2023] [↑](#footnote-ref-4)