RESSOURCES PARTAGÉES

La chronique « ressources partagées » vise à mettre à disposition du plus grand nombre des ressources pédagogiques sur l'enseignement de la statistique à tout niveau. Prendront place naturellement dans cette rubrique des descriptifs de séances de travaux pratiques permettant d'introduire ou de consolider une notion particulière, des jeux de données pertinents pour aider à la visualisation et la compréhension de tel ou tel concept, des exercices ou études de cas complets… Toute suggestion peut être faite auprès d'Antoine Rolland[1], responsable de cette chronique.

Pour commencer cette série nous avons fait appel à Ricco Rakotomalala, de l'Université Lumière Lyon II, afin qu'il nous présente un de ses cours disponibles sur son site http://ricco-rakotomalala.blogspot.fr/ .

Nous avons choisi le cours d'introduction à l'apprentissage supervisé. Il nous semble que le diaporama présenté ici permet de bien cibler les notions fondamentales à enseigner dans une telle séance d'introduction, tout en offrant de nombreuses pistes de développements possibles. En particulier, la ressource présentée ici permet d'aborder dans un format court les notions suivantes : types de variable, principe de la classification supervisée, classifieur bayésien, erreurs de prédiction, matrice de confusion, ensemble d'apprentissage et ensemble de test. Une description précise de chaque diapositive permet de comprendre quels sont les prolongements possibles à présenter aux étudiants pour que ce cours d'introduction donne envie d'en apprendre davantage.

Enseignant chercheur en poste à l'Université Lumière Lyon 2, Ricco Rakotomalala enseigne l'informatique, la statistique et le data mining en Licence et Master. Il est responsable du Master SISE (Statistique et Informatique pour la Science des donnéEs), formation en science des données. Il a développé et mis en ligne des logiciels gratuits et ouverts (*open source*) de fouille de données (*data mining*) depuis une vingtaine d'années, les plus connus étant Sipina V3 et Tanagra. Ces dernières années, il consacre beaucoup d'effort à l'écriture et la diffusion d'ouvrages, de supports de cours et de tutoriels libres, relatifs à ses domaines de prédilection. Les didacticiels décrivent l'utilisation des outils qu'il a développés, mais également d'autres outils tels que R, Python et leurs paquetages spécialisés.

[1] Université Lumière Lyon 2, antoine.rolland@univ-lyon2.fr

INTRODUCTION À L'APPRENTISSAGE SUPERVISÉ

Ricco RAKOTOMALALA[2]

1 Présentation de la ressource

« Introduction à l'apprentissage supervisé » est un diaporama de présentation des principes de la classification supervisée, que l'on nomme également classement ou discrimination, selon notre obédience (plutôt informatique ou plutôt statistique).

La ressource sert de support à un cours magistral. La durée de présentation est à peu près d'1h15, fluctuant selon le degré de réactivité de l'auditoire. Elle est peu détaillée et se prête à de nombreuses digressions. En fonction de la nature et du degré de qualification de la formation visée, de nombreux rapprochements peuvent être faits avec les notions que les étudiants ont abordées durant leur cursus.

Nous nous inscrivons dans un contexte plus large d'enseignement de l'apprentissage supervisé. Ce cours introductif est utilisé comme un préalable aux techniques telles que les arbres de décision, l'analyse discriminante ou la régression logistique. Il sert de cadre générique à ces méthodes.

Dans ma pratique, ce thème est suivi de la présentation du ciblage marketing (*scoring*). Dans cette optique, au-delà du classement brut, l'accent est mis sur le rôle de la probabilité d'affectation qui permettra de développer le concept de « score » par la suite.

Cette séance de cours est complétée par des travaux pratiques sur machines que nous présenterons plus loin (Section 5).

2 Public

Le thème est précis mais le support est relativement dépouillé. Il peut ainsi être adapté à différents publics. Le discours permet d'en préciser la teneur en fonction de la culture ambiante. En ce qui me concerne, je l'ai utilisé ou je l'utilise actuellement dans les formations suivantes.

- DUT STID (Statistique et Informatique Décisionnelle) 2e année ou Licence Professionnelle Statistique. La matière s'inscrit dans la continuité de leur formation. L'effort porte avant tout sur l'explicitation des spécificités du schéma prédictif lorsque la variable cible est qualitative, tant dans la construction des modèles que lors de leur évaluation.
- Master 1 Informatique. Le public est assez hétérogène dans le master où j'interviens. Certains connaissent bien la statistique, d'autres ont un profil plutôt informatique. Avant d'aborder la technique proprement dite, je fais un travail préalable consistant à redéfinir la démarche propre à la fouille de données (*data mining*), puis à situer les différentes

[2] Université Lumière Lyon 2, ricco.rakotomalala@univ-lyon2.fr)

techniques existantes. Les étudiants doivent pouvoir se positionner sur cet échiquier global, en comptabilisant les techniques qu'ils ont déjà pu étudier, et celles qui pourraient leur permettre de compléter leur formation. Ce préalable étant posé, les étudiants n'ont aucun problème à intégrer les notions techniques.

- Master 2 Econométrie. Les étudiants connaissent parfaitement la statistique et l'économétrie, mais pensent ne rien savoir de la fouille de données (*data mining*), ne parlons même pas de la « science des données (*data science*) » qui leur paraît particulièrement mystérieuse. Une grande partie de mon travail consiste à faire le parallèle entre les principales notions de fouille de données (*data mining*) et les thèmes qu'ils ont abordés durant leur cursus, que je connais bien pour l'avoir moi-même pratiqué. La régression logistique, qu'ils ont étudiée sous l'angle de l'économétrie des variables qualitatives, constitue un très bon contrepoint pour avancer durant l'exposé.
- Master 2 Business Intelligence et Master 2 Sécurité Informatique en Formation Continue. Le public est composé de professionnels qui reviennent à l'Université acquérir de nouvelles compétences dans le cadre de la formation continue. Les apprenants viennent d'horizons très divers, ils ont pour point commun d'avoir développé une aversion forte aux mathématiques. La solution passe par une simplification des concepts et, surtout, par de nombreuses applications directes sur machine pour qu'ils puissent appréhender concrètement les notions abordées (par exemple : probabilité = proportion, probabilité conditionnelle = fréquence conditionnelle, …).

Dans tous les cas, plusieurs notions clés doivent être assimilées à l'issue du cours. Elles sont résumées dans la section suivante.

3 Objectifs de la ressource

Le support a pour objectif d'introduire les principales notions de la classification supervisée, c'est-à-dire un schéma prédictif où la variable cible est qualitative. Deux aspects importants sont abordés : la modélisation, la construction du modèle prédictif à partir d'un échantillon de données ; l'évaluation des performances via un ou plusieurs critères numériques.

Pour la première partie, nous nous appuyons sur la théorie de la décision bayésienne, en occultant dans un premier temps les notions de coûts.

Pour la seconde, nous mettons en avant la matrice de confusion confrontant les classes prévues et observées. Plusieurs indicateurs, répondant à différentes interprétations possibles des performances, sont présentés. La notion de coût de mauvaise affectation est introduite à ce stade, uniquement sous l'angle de l'évaluation *a posteriori* des modèles.

4 Présentation détaillée de la ressource

Dans cette section, chaque page du support – à l'exception de la page de titre – est décrite de manière détaillée en mettant en avant les objectifs, le contenu, et le discours que je tiens face aux étudiants, qui peut varier selon les formations.

4.1 Tableau de données (Diapositive 2)

Nous travaillons à partir de données : cette diapositive précise les termes techniques relatifs au tableau « individu – variables » couramment utilisé en statistique (FIGURE *1*). Préciser le vocabulaire est important, en particulier auprès des étudiants ayant une culture statistique peu affirmée.

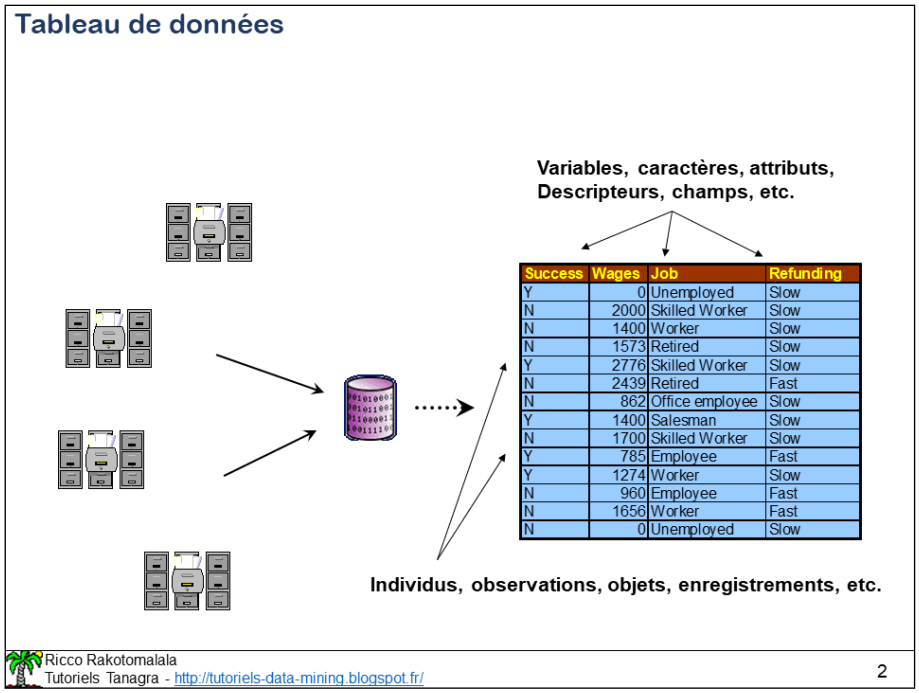

FIGURE 1 – *Tableau de données (Diapositive n°2)*

Pour ces apprenants justement, je prends le temps de distinguer les différents types de variables : qualitative nominale, qualitative ordinale et quantitative. Pour les distinguer, je leur conseille d'essayer de calculer la différence entre les valeurs. L'opération est possible et a du sens pour les variables quantitatives, ce n'est pas le cas pour les nominales. Il y a débat en revanche lorsqu'elles sont ordinales. La question du codage est alors évoquée, notamment la légitimité du codage $\{0, 1, 2, ...\}$ des modalités des variables qualitatives ordinales.

Pour toutes les formations, la question de la transformation de type est également posée. Pour le passage du nominal vers le numérique, le codage disjonctif complet est très facilement identifié par les étudiants. La discussion porte sur le nombre d'indicatrices à générer (autant de variables binaires que le nombre de modalités ?). L'inverse, la discrétisation des variables quantitatives, engendre plus de débats. Il faut choisir le nombre d'intervalles et les bornes de découpage. Les étudiants statisticiens citent quasi-systématiquement les techniques de découpages en intervalles de fréquence ou de largeur égales, mais sans vraiment avoir du recul par rapport aux conséquences que cela peut avoir sur la nature des informations véhiculées par les données (modification des distributions par exemple). Enfin, de nouveau, le traitement des variables ordinales pose problème. Mon objectif est d'amener les étudiants à s'interroger sur les pertes d'informations ou, pire, sur les informations inconsciemment injectées dans les données, lors des opérations de codage ou recodage des variables.

4.2 Statut des variables (Diapositive 3)

Dans la diapositive suivante, le rôle des variables dans la classification supervisée est précisé (FIGURE 2). L'objectif consiste à prévoir les valeurs d'une variable cible Y (variable à prédire, à prévoir, variable endogène, attribut classe, tout dépend de notre culture) à partir d'un ensemble de variables prédictives (X_1, X_2, ...) (descripteurs, variables exogènes, etc.). La variable cible Y est une variable qualitative nominale, souvent binaire ($K = 2$ modalités), mais la situation où $K > 2$ doit pouvoir être gérée également. Les variables prédictives (X_1, X_2, ...) peuvent être de type quelconque, qualitatives ou quantitatives.

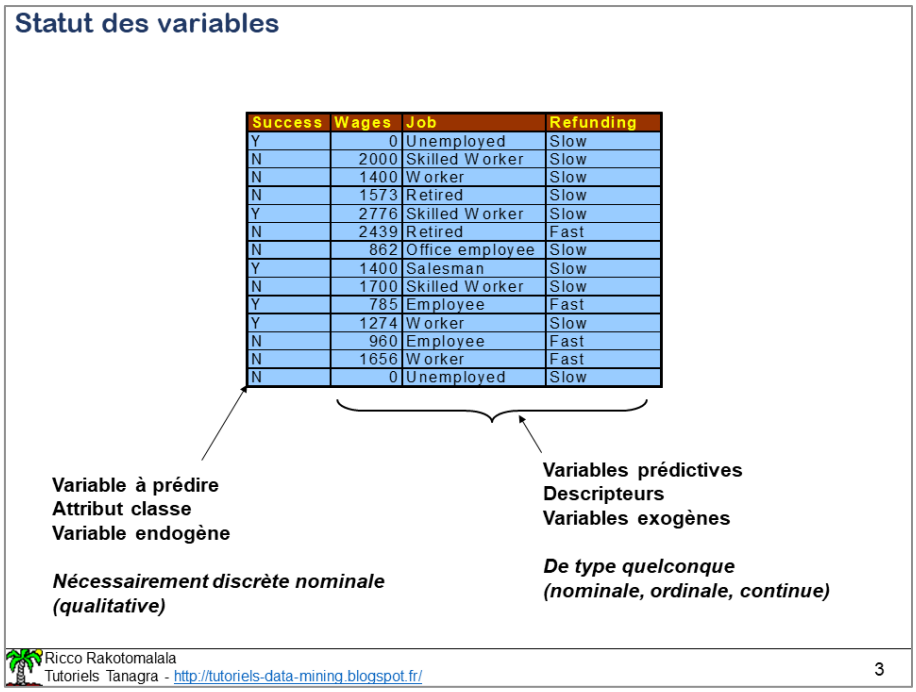

FIGURE 2 – *Statut des variables dans l'analyse (Diapositive n°3)*

4.3 Principe de l'apprentissage supervisé (Diapositive 4)

Cette diapositive est fondamentale. Elle précise le principal objectif de l'apprentissage supervisé : (1) construire un classifieur, un modèle prédictif, (2) qui soit le plus précis possible.

Dans mon discours, le modèle apparaît comme une fonction paramétrée reliant les variables prédictives à la cible. Deux questions clés sont soulevées : quelles sont les différentes formes de liaison fonctionnelle que l'on pourrait utiliser, quels en seraient les paramètres ?

Concernant le choix de représentation, en cohérence avec les approches que je présente par la suite dans mon cours, je mets souvent en avant les systèmes fondés sur des règles et les classifieurs linéaires. Dans le premier cas, les paramètres sont les valeurs de références utilisées dans les conditions des prémisses des règles ; dans le second, il s'agit des coefficients de la combinaison linéaire associée aux descripteurs. Selon la trajectoire antérieure des étudiants, j'entame alors une discussion concernant les différentes méthodes qui pourraient correspondre à ces systèmes de représentation. Les économètres en particulier identifient

facilement la régression logistique. Mais le terme de classifieur « linéaire » les perturbe. On leur a toujours présenté l'approche comme une régression « non-linéaire ». Les échanges à ce niveau sont particulièrement intéressants.

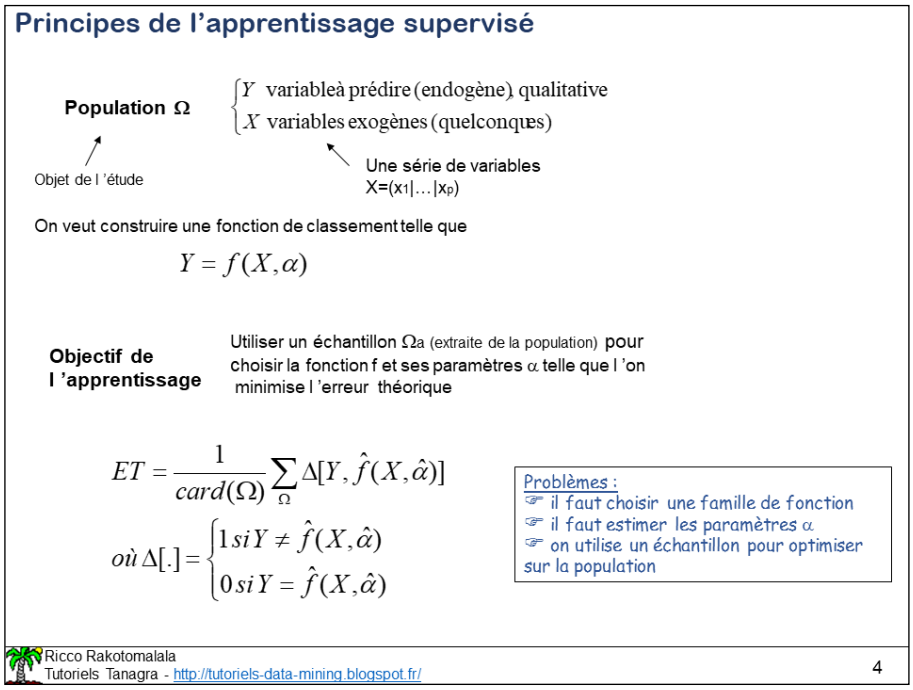

FIGURE 3 – *Principes de l'apprentissage supervisé (Diapositive n°4)*

Modéliser est une bonne chose, mais il faut disposer d'un critère numérique pour mesurer la qualité du modèle prédictif. La seconde partie de la diapositive s'y attelle. Nous occultons les notions de coûts de mauvais classement pour présenter l'erreur théorique qui correspond simplement au rapport entre le nombre de mal classés et le nombre total des individus. L'erreur théorique est ainsi une proportion définie sur l'ensemble de la population. Les étudiants identifient facilement la notion de probabilité de mauvais classement dont ils comprennent le sens. Le domaine de définition est connu [0, 1]. Ils comprennent également qu'avec un modèle parfait, la probabilité de se tromper serait nulle. Plus difficile en revanche est l'identification du pire modèle. Une erreur théorique à 1 leur paraît exagérée, la valeur 0.5 est souvent avancée pour un problème à $K = 2$ classes. Pour $K > 2$, un grand silence s'installe souvent. J'avance l'idée de « classifieur par défaut » sans dire réellement ce qu'il en est. Je réserve la discussion pour les travaux dirigés.

4.4 Cadre probabiliste – Modèle bayésien ($K = 2$ classes) (Diapositive 5)

Cette diapositive (FIGURE 4) est cruciale car elle établit un prisme que j'utilise systématiquement dans la suite de mon cours, lorsque j'aurai à présenter les arbres de décision et l'analyse discriminante linéaire (par exemple en Master 1 Informatique).

Le modèle bayésien assure la minimisation de l'erreur théorique. Je mets l'accent sur deux éléments très importants : (1) la méthode repose sur une estimation des probabilités conditionnelles, (2) mais en réalité c'est le mode qui importe (qui devrait suffire) lorsque nous devons classer les individus. Selon les trajectoires des étudiants (des formations), des

échanges très intéressants peuvent avoir lieu. L'idée est de les emmener à identifier ces deux aspects dans les méthodes prédictives qu'ils ont pu étudier par ailleurs durant leur cursus. Malgré les années d'enseignement, je constate toujours avec un peu de surprise que des étudiants, pourtant parfaitement opérationnels sur certaines techniques supervisées (mise en œuvre, paramétrage, utilisation des logiciels), ont du mal à prendre du recul et sont déroutés par ces questions.

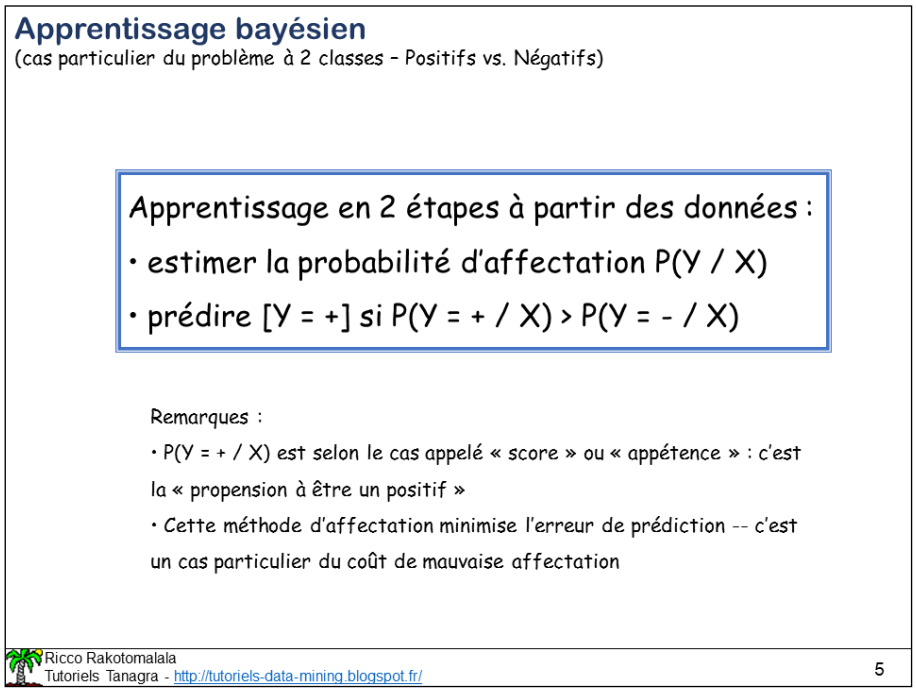

FIGURE 4 – *Modèle bayésien - $K = 2$ classes (Diapositive n°5)*

Dans un cadre de discrimination binaire ($K = 2$), une modalité dite « positive » est souvent plus importante que l'autre (la modalité « négative ») dans les études : nous souhaitons identifier les personnes appétentes à telle ou telle promotion, celles qui risque de présenter telle ou telle maladie, celles qui vont faire des infidélités à leur opérateur téléphonique, etc. Je parle alors de la pratique du « scoring », une application phare du datamining. La notion de « score » est mise en relation avec la probabilité conditionnelle $P(Y = +/X)$.

4.5 Généralisation à K classes ($K \geq 2$) (Diapositive 6)

La généralisation à $K > 2$ ne pose aucun problème dans l'esprit des étudiants (FIGURE 5). Etonnamment, nombreux sont les étudiants qui ne connaissent pas l'opérateur « argument » (arg). Je veille à ne pas aller trop vite dans la présentation.

Dans certaines formations, commencer à parler de calcul probabiliste engendre très rapidement un ennui (accablement) profond auprès des étudiants. Il faut absolument être concret, c'est le rôle de la diapositive suivante.

Pour l'heure, si la formation s'y prête, j'essaie de mettre en relation ce discours avec la notion de système de représentation des modèles prédictifs développé dans la diapositive n°4. Un système de règles correspond au produit cartésien des variables prédictives, pourvu

qu'elles soient toutes qualitatives nominales. Si je m'y prends bien, les étudiants doivent facilement identifier l'écueil que représente la croissance exponentielle du nombre de règles qu'engendre l'augmentation du nombre de descripteurs avec ce dispositif. Une réflexion autour de la question de la sélection de variables pertinentes pour la prédiction est avancée, sans aller trop loin à ce stade.

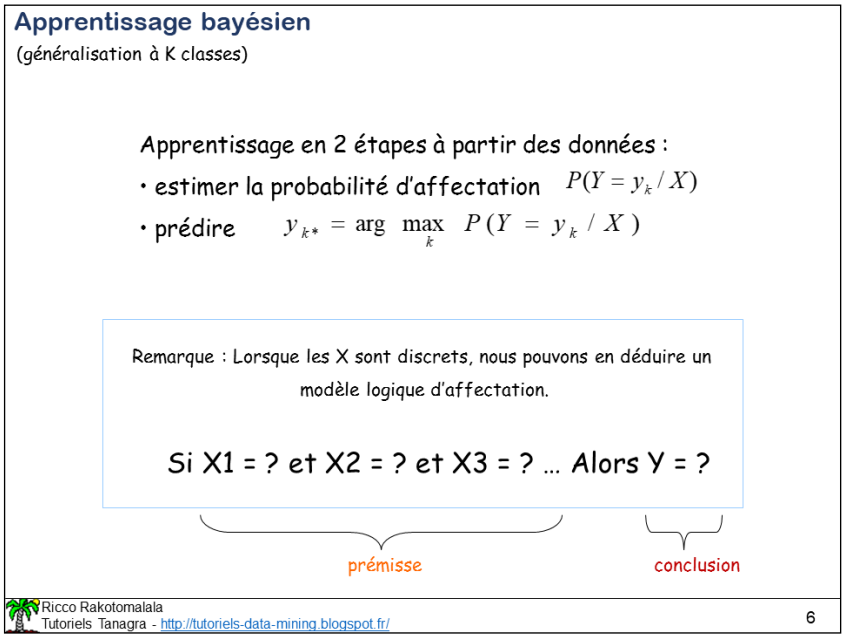

FIGURE 5 – *Modèle bayésien - Généralisation à $K > 2$ classes (Diaporama n°6)*

4.6 Un petit exemple(Diapositive 7)

Les notions de probabilités, surtout conditionnelles, peuvent paraître abstraites. Une application sur un jeu de données réduit permet de préciser les idées, l'exemple est loufoque à dessein pour éveiller l'intérêt de l'auditoire (FIGURE 6).

Les étudiants s'interrogent parfois sur la légitimité de travailler sur un aussi petit fichier ($n = 10$ observations) avec des spécifications particulièrement simples alors que « tout le monde » parle de *big data*. Il faut bien insister sur la vocation pédagogique de ces données, elles permettent de se poser les bonnes questions quant aux problématiques rencontrées en classification supervisée.

Dans un premier temps, je leur demande de construire le modèle permettant de prédire l'occurrence de la « Maladie » (« présent » = modalité positive). A ce stade déjà, la question de sélection des variables pertinentes est posée. Nous disposons de 4 variables candidates. Laquelle ou lesquelles devons-nous choisir ? N'appartient-il pas plutôt à la technique de sélectionner « automatiquement » les bonnes variables ? Ou tout du moins de fournir des indications sur leur influence ?

Je leur propose alors de s'en tenir à la variable « taille » dans un premier temps. Nous calculons les fréquences conditionnelles de « Maladie » en fonction de « Taille », puis nous en déduisons les règles prédictives conformément au modèle bayésien. Les règles sont :

1. **Si** Taille = Trapue **Alors** Maladie = Absente
2. **Si** Taille = Elancée **Alors** Maladie = Présente

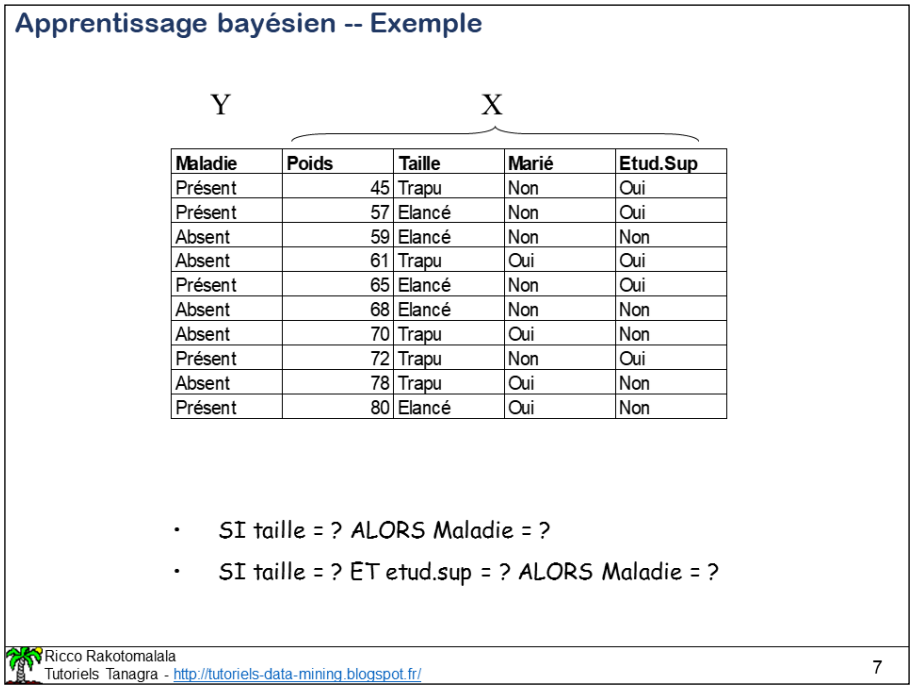

FIGURE 6 – *Un jeu de données exemple (Diapositive n°7)*

Des indicateurs de qualification des règles sont avancés à partir du calcul des fréquences conditionnelles : le support de la règle, comme indicateur de fiabilité de la règle (elle est restreinte au support de la prémisse dans le cadre prédictif) d'une part ; la confiance, comme indicateur de précision de la règle d'autre part. Sur la première règle, nous avons respectivement 5 (en valeur absolue) et 3/5 = 60%. A ce stade, la notion de spécificité des règles devrait germer dans l'esprit des étudiants. On peut améliorer la confiance des règles en ajoutant de l'information (des variables prédictives) dans la prise de décision. Je propose alors de raffiner la première règle à l'aide de « études supérieures ».

La règle « **Si** Taille = Trapue **Et** Etudes Supérieures = Oui **Alors** Maladie = Présente » est identifiée, avec une confiance de 2/3 = 66% mais, a contrario, un support de 3 individus. La règle s'avère plus précise, mais sa portée est moindre, elle est moins généralisable.

L'objectif de l'exemple est de faire réfléchir les étudiants sur les concepts de généralité et spécificité des modèles prédictifs. Et de mettre en rapport cet arbitrage avec la complexité des modèles, qui peut s'exprimer notamment par le nombre de descripteurs que l'on intègre dans la prévision. La question de la sélection des variables pertinentes est latente encore une fois.

Enfin, le traitement de « poids » qui est une variable quantitative est abordé. Dans le modèle bayésien multinomial, l'utilisation directe des valeurs pour définir les règles n'est pas possible car aboutirait justement à des règles trop spécifiques. Les étudiants comprennent bien la notion à ce stade. Les discussions permettent souvent de mettre en évidence la solution de la discrétisation des variables quantitatives pour améliorer la généralité. Avec bien sûr les questions importantes que constituent le choix du nombre d'intervalles et des bornes de découpage déjà abordées dans la diapositive n°2.

Au final, nous passons beaucoup de temps sur cette diapositive qui s'avère cruciale dans la compréhension de la modélisation prédictive. Les étudiants doivent (si possible) avoir en tête les principales questions qu'ils devront se poser lorsque nous embrayerons sur les

« vraies » méthodes de classification telles que les arbres de décision, l'analyse discriminante ou d'autres. Est-ce que la technique étudiée intègre un mécanisme de sélection de variables ? A quel niveau se situe l'arbitrage entre la généralité et la spécificité ? La question du recodage préalable de certaines variables se pose-t-elle ?

4.7 Bilan – Avantages et inconvénients de l'approche (Diapositive 8)

Le bilan s'appuie sur les écueils soulevés lors du traitement de l'exemple jouet. Cette approche directe est certes optimale au sens du taux d'erreur, mais elle n'est pas utilisable dans la pratique (FIGURE 7). Sans nécessairement trop développer puisque ces thèmes seront étudiés par la suite dans mon cours, je parle des méthodes qui s'appuient sur différents artifices pour estimer la probabilité d'affectation ou son mode, soit par des restrictions dans l'espace de recherche (par exemple, les arbres de décision avec la stratégie « divide and conquer »), soit en introduisant des hypothèses statistiques permettant de réaliser les calculs (par exemple, l'analyse discriminante linéaire avec l'hypothèse de normalité des distributions conditionnelles).

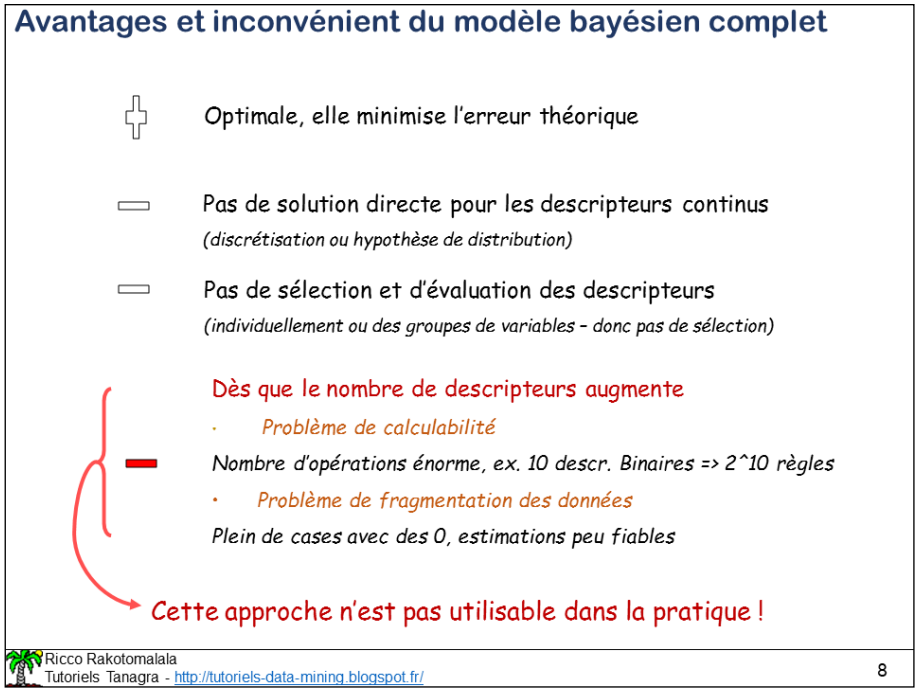

FIGURE 7 – *Avantages et inconvénients de l'approche présentée (Diapositive n° 8)*

4.8 Evaluation de l'apprentissage (Diapositive 9)

L'évaluation des modèles constitue l'étape suivante (FIGURE 8). La performance prédictive sera détaillée plus loin. Dans cette diapositive, nous mettons l'accent sur les aspects qualitatifs : la compréhensibilité des modèles d'une part, leur rapidité et capacité à traiter des grandes bases d'autre part.

Le second thème paraît évident dans le contexte actuel des données massives (*big data*). Les étudiants comprennent très bien que l'aptitude à traiter des grandes bases est un enjeu fort qui dépend à la fois des capacités des machines, de l'organisation des calculs (calcul distribué

avec les technologies Hadoop ou Spark), et des caractéristiques des algorithmes d'apprentissage.

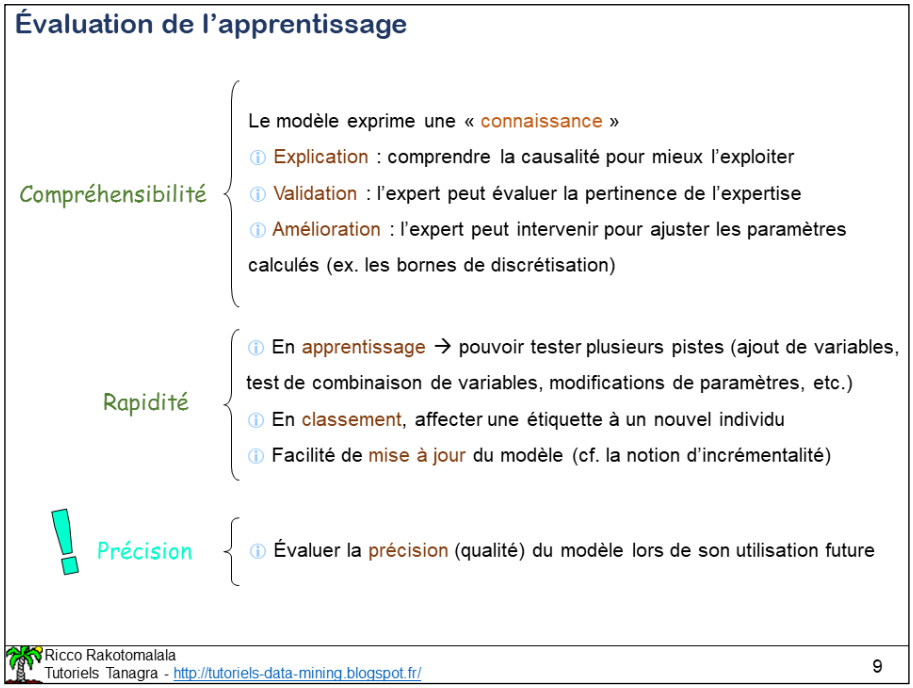

FIGURE 8 – *Evaluation des modèles (Diapositive n°9)*

Le premier thème mérite qu'on s'y attarde un peu plus. Dans de très nombreux domaines, au-delà des prouesses numériques pures, nous avons besoin de comprendre le mécanisme d'affectation aux classes, et en particulier le rôle que jouent les variables. L'idée mise en avant est « un décideur n'adhère que s'il comprend ce qu'il manipule ». Les échanges avec les étudiants s'articulent alors sur l'identification des domaines où cette compréhension est déterminante, et ceux où elle ne le serait pas.

4.9 Matrice de confusion – Cas de $K = 2$ classes (Diapositive 10)

Après ce préambule, nous rentrons dans l'évaluation numérique en remémorant la notion d'erreur théorique mise en avant lors de la présentation de la classification supervisée (Diapositive n°2).

Nous travaillons sur un échantillon et non plus sur la population cette fois-ci. Pour rendre concret le propos, surtout pour les formations où la culture statistique est peu développée, je reviens sur l'exemple de la diapositive n°7 où, à partir du modèle prédictif basé uniquement sur la « taille », je construis au tableau la colonne « prédiction ». En l'opposant à la colonne « Maladie », je présente la matrice de confusion (FIGURE 9).

Je m'en tiens à la discrimination binaire ($K = 2$) dans mon propos. Le passage à K ($K > 2$) classes est effectué lors des travaux dirigés. Différentes notions liées à la lecture de la matrice de confusion sont abordées. L'association entre le taux d'erreur observé (empirique) et l'erreur théorique est facilement identifiée par les étudiants.

En distinguant la modalité d'intérêt (modalité positive de la variable cible), le déchiffrage de la matrice est enrichi avec les notions de « sensibilité » et de « précision » notamment, qui

propose une interprétation élargie du comportement des modèles. Le principal message est : la quantité globale de l'erreur est un critère important, nous l'avons avec le taux d'erreur, mais la structure de l'erreur l'est tout autant, les autres ratios permettent d'en rendre compte.

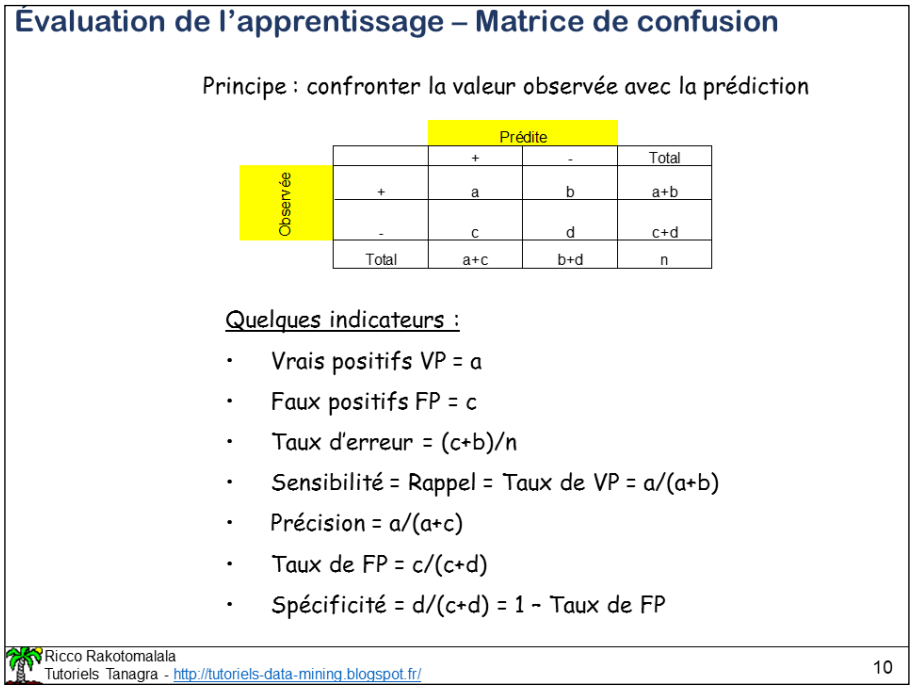

FIGURE 9 – *Matrice de confusion (Diapositive n°10)*

Prendre des illustrations de la vie courante est appréciable pour rendre le discours moins abstrait. Elles sont nombreuses. Un article concernant la fraude aux allocations paru dans le journal « Le Progrès » (nous sommes à Lyon) datant de 2013 fait partie de mes exemples favoris[3]. Il s'agit d'un journal régional généraliste et pourtant on y parle de fouille de données (*data mining*), et plutôt de manière technique si l'on se penche attentivement sur l'article. L'exercice consiste à y déceler la matrice de confusion qui est décrite de manière non explicite, et d'identifier/reconnaître les ratios qui nous intéressent.

4.10 Quantité vs. Structure de l'erreur – Coût de mauvaise affectation (Diapositive 11)

Pour approfondir la notion de « structure de l'erreur », nous traitons un exemple où deux modèles aboutissent à des matrices de confusion différentes (FIGURE 9). L'objectif est de désigner le « meilleur » modèle au sens des critères numériques.

Les deux classifieurs proposent le même taux d'erreur. Ce serait trop facile sinon. L'enjeu consiste alors à calculer les autres indicateurs et à définir celui qui serait le plus intéressant. On devine bien qu'il n'y a pas de réponse univoque à cette question. Tout l'intérêt ici est de pousser les étudiants à réfléchir en termes de cahier des charges et, selon les attentes du maître d'ouvrage, d'identifier les critères les plus adaptés compte tenu du contexte.

[3] « Frauder les allocs, un 'sport' de plus en plus risqué », Journal Le Progrès, 30.01.2013 ; (accessible en ligne) http://www.leprogres.fr/france-monde/2013/01/30/frauder-les-allocs-un-sport-de-plus-en-plus-risque

Cela m'amène à parler des coûts de mauvaise affectation. Je rappelle que nous simplifions très souvent en considérant qu'elle est unitaire et symétrique, c.-à-d. un mauvais classement coûte 1, un bon classement vaut 0. C'est rarement le cas dans les études réelles. Les conséquences de la prédiction de la présence d'une maladie chez une personne saine sont très différentes de la situation inverse. J'essaie de faire passer deux messages clés à ce propos : la définition de la matrice de coûts est très difficile (unités, valeurs), elle ne nous revient pas en tant que *data miner* ; mais nous pouvons calculer un indicateur étendu, le coût moyen de mauvais classement, si les décideurs nous fournissent cette matrice, en la croisant avec la matrice de confusion. Le jeu consiste alors à demander aux étudiants d'identifier la situation (la matrice de coûts) où ce nouvel indicateur serait équivalent au taux d'erreur.

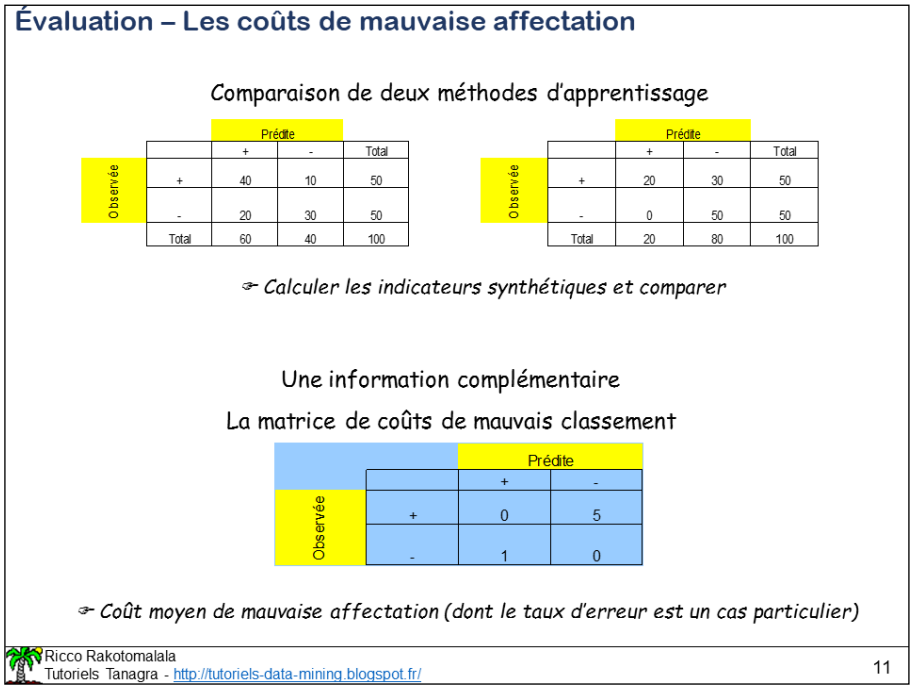

FIGURE 10 – *Comparaison de modèles – Introduction de la matrice de coûts (Diaporama n°11)*

4.11 Schéma apprentissage-test pour l'évaluation (Diapositive 12)

Enfin, *last but not least*, l'avant-dernière diapositive développe l'idée selon laquelle les mêmes données ne peuvent pas être partie (servir à la constitution des modèles) et juge (servir à leur évaluation) (FIGURE 11). Le principe en est facilement admis par l'auditoire, sauf lorsqu'il s'agit d'économètres – je reviens dessus plus bas (Section 6) –, surtout lorsque j'explique qu'il suffirait de définir une règle par individu pour être assuré d'obtenir un taux d'erreur nul si on utilise les mêmes données pour l'évaluation (à condition que l'étiquette ne soit pas bruitée, mais il vaut mieux ne pas rentrer dans ce genre de considérations pour ne pas perdre les étudiants, du moins à ce stade). On pourrait également développer un peu plus en parlant de surapprentissage si le public s'y prête.

La solution passe par la subdivision aléatoire des données en échantillons d'apprentissage et de test avec approximativement des proportions à 70% et 30%, ou encore 2/3 et 1/3. Dans cette phase d'initiation, il ne me paraît pas opportun d'approfondir la légitimité de ces

proportions. Je dis simplement qu'il s'agit de la pratique usuelle, qu'elles correspondent au paramétrage par défaut dans la majorité des logiciels.

Au tableau, je complète mon discours en retraçant dans sa globalité la démarche de classification supervisée en faisant apparaître explicitement les différentes phases (FIGURE 12).

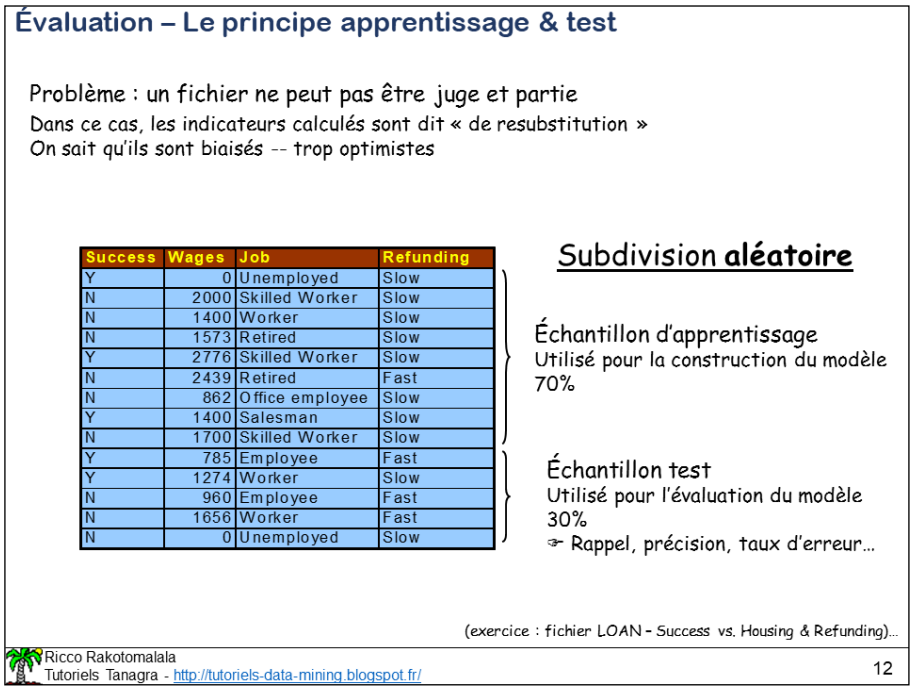

FIGURE 11 – *Subdivision des données en échantillons d'apprentissage et de test (Diapositive n°12)*

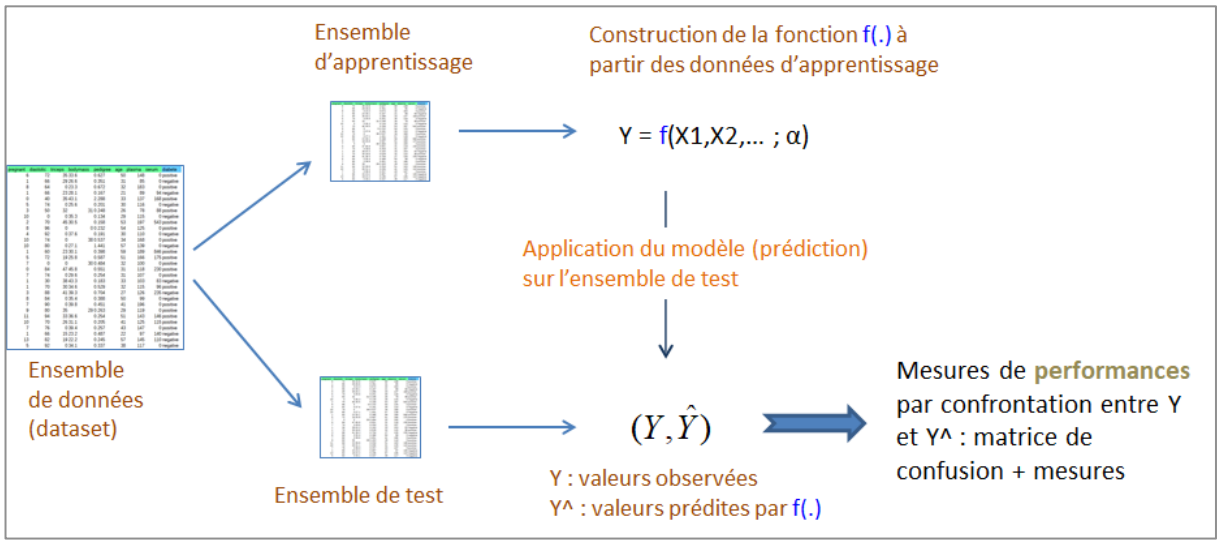

FIGURE 12 – *Différentes phases de la classification supervisée*

4.12 Bibliographie (Diapositive 13)

Je me contente de deux références dans la bibliographie. Il s'agit de deux excellents ouvrages qui correspondent à l'époque où j'avais écrit la première version de ce diaporama, à savoir :

- Bardos M. (2001), *Analyse discriminante – Application au risque et scoring financier*, Dunod ;
- Hastie T., R. Tibshirani, and J. Friedman (2001), *The elements of statistical learning – Data mining, Inference and Prediction*, Springer.

Au fil des années, je passe rapidement sur cette partie pour plutôt réaliser une recherche Internet avec les étudiants en entrant les mots clés en relation avec le thème du jour. J'espère ainsi les pousser à compléter par eux-mêmes leurs connaissances. L'exercice n'est pas facile parce qu'il y a une profusion d'information difficile à démêler sur le web. Disposer de quelques repères utiles peut les aider, notamment lorsque l'on utilise des mots clés en anglais.

5 Compléments de la ressource

Ce cours est complété par des travaux dirigés, sous Excel toujours dans un premier temps, puis sous R ou Python selon les formations. L'objectif est de retracer sur machine les différentes phases exposées dans le diaporama : subdivision des données en échantillon d'apprentissage et test, je le fais pour les premiers exercices, puis les étudiants s'en occupent eux-mêmes par la suite ; calcul des probabilités conditionnelles sur l'échantillon d'apprentissage avec des indications sur les variables à utiliser ; déduction des règles prédictives ; prédiction sur l'échantillon test ; et, enfin, calcul de la matrice de confusion et des indicateurs de performances.

Mes données de prédilection proviennent du serveur « UCI – Machine Learning Repository » (https://archive.ics.uci.edu/ml/datasets.html). J'utilise en particulier les bases : « Votes au congrès » [*Congressional Voting Records Data Set*] et « Maladies cardiaques » [*Statlog (Heart) Data Set*]. Cette dernière comporte des descripteurs quantitatifs qui me permettent d'introduire la discrétisation dans les exercices.

6 Retour d'expérience

C'est un enseignement que j'assure depuis de nombreuses années. J'en cerne à peu près les contours.

Les étudiants en statistique et, plus généralement, les apprenants en continuité d'études ayant de bonnes notions de probabilité n'ont aucune difficulté pour intégrer les principes de l'apprentissage supervisé. Le schéma global est rapidement assimilé. Les travaux pratiques sur machine se déroulent très bien généralement.

Les économètres sont parfois un peu déroutés par l'utilisation d'un échantillon distinct pour l'évaluation des performances en prédiction. Cette pratique ne correspond pas vraiment à l'usage en économétrie. Je me réfère souvent au critère Akaike (AIC) pour expliciter le dispositif. Je m'appuie sur l'analogie entre l'évolution du taux d'erreur en test et celle du critère AIC en fonction de la complexité des modèles. Lorsque ces derniers reposent sur des

systèmes de représentation différents, le nombre de paramètres n'est plus comparable, la nécessité d'un arbitre impartial sous la forme d'un échantillon n'ayant pas participé à l'apprentissage paraît alors pertinente dans l'esprit des étudiants. J'en profite pour parler de la malédiction de la dimensionnalité, et placer les fameuses courbes d'évolution des taux d'erreur en apprentissage et en test en fonction de la complexité des modèles.

Le problème est tout autre en formation continue. Les apprenants, pour la plupart des informaticiens en ce qui concerne les diplômes où j'interviens, sont dans le monde professionnel depuis bon nombre d'années. Ils conservent souvent un souvenir douloureux des mathématiques et de la statistique. Ma première mission consiste à les rassurer en étant le plus pragmatique possible. A cet effet, je multiplie les exemples en les faisant travailler sur machine, utilisant intensivement le tableur Excel que tous savent, plus ou moins, manipuler. Même si nous travaillons sous R ou Python par la suite, je les fais toujours débuter sous Excel. Les étudiants ne peuvent pas entrer des commandes sans comprendre dans un tableur. Ce point d'entrée simple est important pour qu'ils appréhendent concrètement les notions qui peuvent nous paraître évidentes (fréquence relative, fréquence conditionnelle). La séance a lieu en salle informatique. Nous essayons d'avancer de concert pour que les concepts puissent être traduits directement en calculs et résultats numériques avant la fin de la séance (pour éviter qu'ils partent sur une mauvaise impression). Exposé et travaux pratiques s'inscrivent dans un créneau de 3 heures dans ce cas.