Note du 11 juillet 2017

Voici quelques questions importantes qui me viennent à l’esprit dans la première semaine, qui peuvent m’aider à mieux comprendre l’intention de votre projet et qui mèneront la suite de mon stage et affecteront les méthodes statistiques empruntées. Avec plaisir de les discuter avec vous la prochaine fois.

1. Est-ce que le but de cet examen psychomoteur et la collecte des données durant les années passées s’arrête à hiérarchiser la sensibilité des 26 critères et d’étudier leur importance, ou est-ce qu’en plus de cela, la mise en place d’un barème psychomoteur afin de prédire dans le futur si une personne appartient à un sous-groupe pathologique, uniquement à partir des résultats de ses tests psychomoteurs, est aussi envisagée, voire l’intention la plus importante du projet ? Puisqu’en fonction de ce que l’on veut faire, les méthodes statistiques appliquées peuvent se différer. En effet, dans le cas où la prédiction fait partie de l’objectif de la recherche, je proposerais des méthodes de l’apprentissage supervisé (régression logistique, SVM, forêts aléatoires, kNN, etc.), qui sont assez efficaces dans un problème de classification typique comme le nôtre. On pourrait en discuter ensemble et je pourrais éventuellement faire une petite présentation sur ce genre de méthodes.
2. Toujours dans le cas de la prédiction, est-ce qu’une simplification de critères (diminution du nombre de tests psychomoteurs) est également envisagée ? Si le passage de tous les tests est coûteux ou compliqué, ou encore si l’on veut rendre le résultat de l’examen plus explicable en gardant seulement les tests les plus représentatifs, c’est bien la situation dont on parle. Dans ce cas-là des traitements visant à limiter le nombre de critères peuvent être considérés. En revanche, si le passage et la conservation de tous vos tests ne sont pas compliqués et c’est bien la qualité de la prédiction qui est au premier rang, alors ce n’est pas forcément la peine de faire ça.
3. Comme vous l’aviez certainement aussi remarqué, en général les 26 critères sont de telle façon que souvent pour un certain critère peu de témoins ont obtenu de 0, mais en revanche le nombre de patients avec un résultat 1 est considérable. Cela veut dire que la plupart des critères sont « de mauvaise qualité » en termes de séparation des deux groupes. On aurait probablement du mal à différencier parfaitement les deux catégories en choisissant seulement 3, 4 critères les plus sensibles. De l’autre part, toutefois, si on ne considère que le nombre de témoins qui ont raté un test comme seule mesure de qualité, la plupart des critères sont quand-même très bons. Cela veut dire qu’on ne parle ici plus de la séparation totale, mais d’une « séparation au sens unique » : si une personne a obtenu un 0 pour un critère, on sera « quasiment » sûr qu’il a une grande chance de subir une pathologie en question ; par contre s’il a obtenu un 0, on peut rien y dire. La question qui se pose ici est, si dans la pratique le faux positif (la personne qui se fait diagnostiquer à tort une maladie psychiatre mais ne l’a en fait pas) est plus grave que le faux négatif (la personne qui passe l’examen mais possède en revanche cette maladie), ou si c’est l’inverse, ou c’est également important ? ça aura un impact non seulement sur les méthodes de prédiction, mais déjà sur la hiérarchisation des critères.
4. L’étude principale dans mon stage doit d’après vous se trouver notamment dans le champ Patients vs. Témoins, Schizophrènes vs. Témoins ou Dépressifs vs. Témoins ? Ou encore dans tous les trois si le temps nous permet ? Et la détection des sous-catégories de la schizophrénie ça vous intéresserait peut-être aussi (cela exige d’ailleurs plus de données de patients à mon avis) ?
5. En ce qui concerne la cotation binaire : la cotation binaire me gêne en effet pas du tout d’après ma première constatation de vos données. C’est juste que pour faire des analyses statistiques un peu plus avancées sur d’autres sujets intéressants, il vaut mieux avoir plus de données (surtout patients). J’ai cru au début que vous avez tronqué vos données collectionnées durant des années pour vous débarrasser du biais par rapport à l’âge et au sexe. Du coup il ne reste que 71 patients tous en dessous de 40 ans. Mais je viens de me rendre compte que ces 71 patients sont un nouveau jeu de données qui n’ont rien à faire avec l’ancien jeu de données avec lequel vous aviez effectué vos recherches en 2008 et en 2012 ? Et que ce n’est pas facile de transformer les résultats d’une cotation de 0 à 4 en binaire ? Dommage qu’on ne puisse pas profiter des anciennes données.
6. La cotation à 5 niveaux donnera à nos données sans doute une forme plus riche d’un point de vu statistique. De l’autre part je comprends tout à fait la problématique que vous m’aviez expliquée, que cette cotation sera trop subjective pour que l’étude puisse facilement se généraliser. Si jamais vous y hésitez, j’ai un conseil à vous proposer, c’est qu’on ne transmettrait en fin de compte pas un barème concret ou une formule précise pour qu’un médecin puisse déterminer tout de suite la pathologie de la personne en lisant le résultat des tests de celui-ci, mais plutôt une méthode de calcul que le médecin doit ensuite appliquer à sa propre base de données qu’il a établie au fils de temps, afin que la subjectivité ne joue plus aucun rôle car c’est toujours le même médecin. Cela étant dit, c’est vous qui êtes dans la profession et connaissez l’objectif de votre projet beaucoup mieux que moi, je suis donc à votre écoute à cet effet.