Détection de diabètes chez les femmes Amérindiennes du peuple Pima
CR - TP Noté

Présentation du dataset

Lors de ce TP noté, nous nous sommes intéressés au dataset portant sur des cas de diabètes chez les femmes Amérindienne du peuple Pima. Nous avions à disposition un jeu de données composé des données physiologiques et médicales de 768 femmes. Parmi ces mesures :

- Le nombre de fois où la femme a été enceinte
- La concentration de glucose plasmatique
- La pression sanguine diastolique (mmHg)
- L'épaisseur du pli cutané du triceps (en mm)
- Un mesure de l'insuline après 2 heures (en mu U/ml)
- L'IMC (en kg/m^2)
- Une fonction évaluant les antécédents familiaux de diabètes
- L'age
- Une variable diabetes portant l'information sur la présence ou l'absence de la maladie

Problème

La question qu'il nous a alors été demandé de trancher est la suivante :

Peut-on utiliser ces mesures pour modéliser voire prédire la présence/absence de diabètes chez ces femmes ?

Construire un modèle pertinent

C'est la partie qui m'a pris le plus de temps. Avant de commencer à travailler sur les données en elles-mêmes, il fallait considérer le problème, le comprendre et être de capable de prendre une décision éclairée sur la façon dont on souhaite évoluer notre solution.

Après les cours que nous avons suivi ce semestre, on reconnaît facilement un problème de classification, mais comment saura-t-on in fine que notre modèle est pertinent ? Faut-il optimiser sa précision, sa spécificité ou encore autre chose ?

Pour répondre à cette question, j'ai demandé à Aurélie Enjalbert, une étudiante en 5ème année de médecine à l'université d'Aix Marseille de me parler du diabète, de ces risques, de l'importance de sa détection.

Lors de cet entretien, elle m'a confirmé que chacune des mesures physiologiques et médicales dont nous disposons a son utilité dans notre analyse et qu'il ne semblait pas pertinent d'en écarter une plutôt qu'une autre. À cette étape, je n'avais donc aucune raison d'en discriminer une puisqu'une personne qualifiée dans le domaine médicale m'a suggéré de ne pas le faire.

Concernant le diabète, elle m'a appris qu'il s'agit d'une maladie qui n'est pas nécessairement grave à condition de ne pas la détecter trop tard. D'où l'importance de la détection. Il vaut mieux selon elle détecter le maximum de personnes porteuses, quitte à en détecter trop. Nous voulons donc un indicateur de notre modèle qui nous donne un maximum de Vrai Positif par rapport au nombre de Faux Négatifs. L'indicateur qui convient est donc la spécificité donc la formule est la suivante :

$$\text{Sensibilité} = \text{Vrai Positif} / (\text{Vrai Positif} + \text{Faux Négatifs})$$

C’est cet indicateur auquel nous allons nous intéresser et qui rendra compte de la qualité de notre modèle.

Analyse des données

Remarque : Lors de ce projet, j’ai décidé de travailler avec Python car c’est un langage que je connais bien et sur lequel je n’avais pas encore construit de modèle de machine learning. Après coup, il aurait peut être était plus simple de travailler avec R.

Avoir un dataset utilisable

Avant de commencer notre analyse, il faut s’assurer que nos données sont complètes. On remarque ici que ce n’est pas le cas. En effet, il manque un grand nombre de données pour les mesures insulin et triceps qui ont leur valeur pas défaut à 0, ce qui correspond à une absence de valeur et pas à une valeur nulle.

Il faut donc se méfier de ces deux colonnes, car elles pourraient fausser nos résultats. D’un autre côté, il manque une grande partie de ces données et si on enlève toutes les lignes qui possèdent un 0 dans les colonnes insulin et triceps, il ne nous reste que 394 lignes sur les 768 initiales (i.e. on va travailler avec seulement 51.3 % de nos données).

J’avais alors deux solutions :
- Enlever les lignes qui possèdent la valeur 0 pour insulin et triceps
- Enlever les colonnes insulin et triceps

Pour travailler sur des données propres, j’ai choisi de faire les 2. Je travaille donc avec 39,9 % de mes données initiales, ce qui est peu, mais c’est aussi beaucoup moins biaisé.

Convertir les variables catégorielles en variables numériques

On aurait pu avoir à transformer nos variables catégorielles en variables numériques, mais nous n’avons que des variables numériques donc on peut passer à l’étape suivant.

Standardiser nos données

Pour éviter que les variables qui ont des valeurs très importantes ne prennent le dessus sur les autres, on standardise les valeurs entre 0 et 1.

Splitter en training set et test set

On sépare nos données en 2 sous ensembles. Un premier qui correspond à 80% de nos données et qui va servir à entrainer notre modèle et 20% qui serviront à évoluer le modèle une fois que celui ci sera prêt.

Choix et construction du modèle

Avant de m’arrêter sur un modèle de classifieur scikit-learn, j’en ai comparé plusieurs : LogisticRegression, DecisionTreeClassifier, KNeighborsClassifier, GaussianNB et SVC.

Après quelques tests, c’est la régression logistique qui apparaissait comme la meilleure de part les résultats qu’elle fournissait. J’ai donc opter pour ce type de modèle dans la suite de mon analyse.

J’ai ensuite construit mon modèle, je l’ai entrainé avec mes données de train et je lui ai fait prédire les données de test.

Matrice de confusion

Cette matrice nous donne une idée de la qualité de la classification du modèle. Elle s’organise comme cela :

		Réalité	
		Positif	**Négatif**
Prédiction	**Positif**	TP	FP
	Négatif	FN	TN

Comme je l’ai expliqué plus haut, on veut un TP le plus grand possible comparé au FN pour détecter le maximum de personnes diabétique et on s’intéresse pas à la partie jaune. En effet, si on détecte des personnes non malades comme positives, ce n’est pas très grave dans notre cas.

Analyse et interprétation des résultats

On peut remarquer que d’une exécution à l’autre, la valeur de la sensibilité varie en 0,753 et 0,855. Cela s’explique par la faible quantité des données que nous avons à disposition.

On peut tout de même dire que ce sont des résultats corrects au vu de la quantité de données que nous avons.

Amélioration possible du programme

Si nous avions des données complètes et/ou plus de lignes, notre modèle en profiterait largement et la sensibilité augmenterait alors.

On pourrait aussi essayé d’implémenter des modèles plus complexes qui donnerait peut être des résultats plus précis.

Concernant le tunage des hypermètres, j’ai préféré ne pas y toucher. En effet, le fait de travailler avec un jeu de données de cette taille va induire un trop grand biais et ce n’est pas souhaitable.

Sources

Lien du dépot Github :
https://github.com/thomascormier/TP_PimaIndiansDiabetes_classifier

Consignes et données :
https://plmlab.math.cnrs.fr/gdurif_teaching/polytech_ig5_regression_tutorial/-/tree/master/data

Implémentation du modèle avec Python :
https://medium.com/edureka/machine-learning-classifier-c02fbd8400c9

Comparer les classifiers :
https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html
https://www.youtube.com/watch?v=zo4v7r1l58A&ab_channel=GilbertTanner

Visualisation des données via une analyse trouvée en ligne :
https://machinelearningmastery.com/case-study-predicting-the-onset-of-diabetes-within-five-years-part-1-of-3/

Documentation sur le diabète :
Discussion sur le diabète avec Aurélie Enjalbert, étudiante en 5ème année de médecine de l'université d'Aix-Marseille
https://fr.wikipedia.org/wiki/Diab%C3%A8te_sucr%C3%A9