Méthodologie

Datagotchi est une application ludique et éducative qui engage l'utilisateur dans la conception interactive d'un avatar représentant ses habitudes de vie, et qui permet ensuite de prédire des attitudes et comportements sociaux. À partir de caractéristiques individuelles liées aux habitudes de vie, cette nouvelle version de Datagotchi prédit l'intention de vote des utilisateurs en vue des élections générales québécoises du 3 octobre 2022.

Les données des internautes sont collectées de manière confidentielle sur l'application Datagotchi lorsque les utilisateurs remplissent le questionnaire. Les questions sur les caractéristiques sociodémographiques des utilisateurs, ainsi que sur leurs habitudes de vie, permettent de prédire la probabilité qu'a l'utilisateur de voter pour chacun des cinq partis provinciaux suivants : la Coalition avenir Québec, le Parti libéral du Québec, Québec solidaire, le Parti québécois, et le Parti conservateur du Québec. Tous les partis dont les projections de vote étaient de 5% ou plus avant le déclenchement de la campagne sont considérés dans l'analyse. Nous excluons donc les partis plus en marge tels le Parti vert du Québec (PVQ), Mouvement Québec français (MQF) et le Parti canadien du Québec. Pour l'instant, le pourcentage d'électeurs qui votent pour ces partis est trop faible pour alimenter l'algorithme de prédictions. Toutefois, les modèles prédictifs se raffinent et s'ajustent et fonction des répondants : ces partis pourraient éventuellement être inclus dans Datagotchi s'ils atteignaient une certaine popularité auprès de l'électorat québécois.

Recherches préliminaires et sélection des questions sur les habitudes de vie

La sélection des variables liées aux habitudes de vie constitue en soi un défi théorique et méthodologique puisque les caractéristiques liées au style de vie d'un individu sont pratiquement infinies. La sélection des questions posées sur l'application Datagotchi s'est réalisée en plusieurs étapes. Une méthode de recherche de type '*snowballing*' a été privilégiée afin de produire une recension systématique des écrits sur la mesure des habitudes de vie. Notre conceptualisation des habitudes de vie (ou *lifestyle*) englobe des comportements et caractéristiques observables, par exemple les comportements de consommation ou de loisirs (Van Acer, 2015). Les processus cognitifs abstraits (par exemple les valeurs, les croyances, ou les opinions) sont exclus de notre conceptualisation des habitudes de vie, bien que nous ne sous-estimons pas leur importance potentielle pour expliquer les habitudes de vie en elles-mêmes. Le *snowballing* consiste à utiliser la liste des références d'un article important de la discipline (*backward snowballing*) et/ou d'identifier les articles ayant cité un article important (*forward snowballing*) afin d'identifier de nouveaux articles apparentés ou reliés (Wohlin, 2014). Nous avons privilégié une combinaison de ces deux méthodes. L'un des avantages de cette procédure systématique est qu'elle démarre à partir d'articles pertinents (souvent en utilisant Google Scholar et en regardant le nombre de citations) et les utilise ensuite pour orienter la recherche (Wohlin, 2014).

Une liste de références a ensuite été conçue, incluant ou excluant les articles en fonction de leur pertinence[1]. Pour chaque article retenu, les critères de base suivants ont été inclus dans une base de données: année de publication, nombre de citations, auteur(s) et type (par exemple, article, chapitre de livre, etc.) de publication. Chaque référence s'est aussi fait attribuer une note de pertinence allant de 0 à 10, où une note de 10 signifie les articles les plus pertinents. En utilisant ces critères, un indice de pertinence des textes a été créé, et ramène le score de 0 à 2,5, où 2,5 indique les textes plus pertinents. L'indice de classement pondère par ailleurs les articles les plus récents afin de leur donner plus de poids, en acceptant le postulat que ces articles représentent les avancées les plus actuelles en matière de mesure du mode de vie (voir Fréchet, Savoie et Dufresne, 2020, pour la création d'un indice similaire). Cependant, comme l'indice surpondère également

[1]La pertinence relative d'un article a été déterminée par trois chercheurs qui parcouraient le résumé et les parties les plus importantes de chaque article.

les textes les plus cités, les publications plus vieilles ne sont pas pénalisées si elles sont largement citées (ici, on peut penser aux canons, ou classiques, de la littérature sur le sujet). Cette approche systématique nous a permis de classer les sources canoniques sur la mesure du style de vie (voir les Figures 1 et 2 en Annexe). Cette première phase de recherche a été effectuée en 2021 et a permis de sélectionner les variables finales de la première version de Datagotchi, déployée lors des élections fédérales canadiennes de 2021. À l'exception de quelques questions additionnelles, les mêmes questions sont posées dans cette nouvelle version.

Analyses préliminaires

Pour cette nouvelle édition de Datagotchi, deux sondages pilotes ont été déployés et ont permis de collecter des données préliminaires auprès d'environ 3000 répondants québécois. Chacun de ces sondages contient l'ensemble des questions de Datagotchi, en plus de différentes questions sur des attitudes politiques, et les variables sociodémographiques conventionnelles de la recherche en comportement électoral. Les sondages contiennent aussi le choix de vote auto-rapporté afin d'entraîner le modèle. Le premier sondage a été déployé le 25 juillet 2022 (n= ± 1500), et le deuxième sondage a été déployé le 22 août 2022 (n= ± 1500). Les deux sondages ont été pré-stratifiés afin d'assurer une proportion représentative des répondants en fonction de leur âge, de leur genre, de leur langue et de leur lieu de résidence.

Les données de sondage préliminaires, une fois collectées, ont permis d'identifier le type de modèle prédictif le plus performant afin de prédire l'intention de vote, puis d'entraîner les données afin de créer un algorithme d'apprentissage. La première base de données (juillet 2022) fut d'abord séparée en deux parties: les données d'entraînement (*training set*) et les données de test (*testing set*). Les données furent séparées (les observations furent assignées de façon aléatoire) selon les proportions suivantes: 75% pour les données d'entraînement et 25% pour les données de test. 1) Les données d'entraînement permettent d'entraîner le modèle, et sont ensuite utilisées par l'algorithme d'apprentissage; 2) Les données de test, quant à elles, permettent de mesurer le taux de succès du modèle - ou reconnaître la capacité du modèle à faire une prédiction de l'intention de vote qui corresponde à celle du répondant. Elles permettent alors de mesurer l'erreur du modèle final sur des données sur la base desquelles le modèle n'a pas été développé initialement. Les données de test permettent alors d'évaluer la qualité du modèle.

Pour chaque modèle, les informations suivantes sont extraites des données tests:

Le niveau d'exactitude (*accuracy*), qui indique le pourcentage de bonnes prédictions du modèle, ou la capacité du modèle de prédire les bonnes classes et les vrais positifs et négatifs de la variable d'intérêt. On le calcule de la manière suivante:

$$Exactitude = \frac{Vrai\ \ positif + Vrai\ \ négatif}{Total}$$

Vrai positif: Dans ce cas-ci, un exemple serait un utilisateur qui a l'intention de voter pour la Coalition Avenir Québec (CAQ), et dont le modèle prédit un vote en faveur de la CAQ.

Vrai négatif: Dans ce cas-ci, un exemple serait un utilisateur qui n'a pas l'intention de voter pour la CAQ, et dont le modèle prédit que le vote ne va pas à la CAQ.

Le niveau de rappel (*recall*), quant à lui, se concentre sur les faux négatifs et est calculé de la façon suivante:

$$Rappel = \frac{Vrai\ \ positif}{Vrai\ \ positif + Faux\ \ négatif}$$

Faux négatif: Dans ce cas-ci, un exemple serait un utilisateur qui a l'intention de voter pour la CAQ, alors que le modèle prédit que le vote ne va pas à la CAQ.

Enfin, le niveau de précision (*precision*) se concentre sur les faux positifs et est calculé de la façon suivante:

$$Précision = \frac{Vrai\ \ positif}{Vrai\ \ positif + Faux\ \ positif}$$

Faux positif: Dans ce cas-ci, un exemple serait un utilisateur qui n'a pas l'intention de voter pour la CAQ, alors que la modèle prédit un vote en faveur de la CAQ.

Deux mesures, le rappel et la précision, sont utilisées par la suite afin de vérifier si le modèle est capable de reconnaître ce qui est du bruit de ce qui ne l'est pas à l'intérieur des données tests. Ces deux mesures sont combinées afin d'évaluer la performance moyenne du modèle (appelée F-mesure, ou *F-score*) sur les indicateurs de rappel et de précision. On calcule la F-mesure de la façon suivante:

$$F - mesure = 2 \ X \ \frac{Précision \ X \ Rappel}{Précision \ + \ Rappel}$$

Différents modèles ont été testés et comparés: des modèles logistiques, des forêts d'arbres décisionnels et Boosting de Gradient. Certains modèles d'apprentissage machine (par exemple, des réseaux de neurones) ont été écartés, puisqu'ils sont moins intelligibles, et plus difficilement explicables. Autrement dit, ces modèles ont recours à des méthodes dont le fonctionnement ne permet pas aux humains de comprendre l'intégralité des étapes du processus. Puisque la vulgarisation scientifique et le transfert de connaissances est un objectif central de Datagotchi (et que le tableau de bord interactif doit permettre à l'utilisateur de comprendre les données et l'effet de chaque variable sur le modèle) l'intelligibilité de l'algorithme est un critère essentiel dans la sélection du modèle final. À l'heure actuelle, la régression logistique multinomiale est utilisée derrière l'ensemble des prédictions émises par l'application.

Une fois créé et testé avec les données du premier sondage, l'algorithme de prédiction a été testé sur la deuxième base de données (août 2022). L'utilisation de la deuxième base de données comme *testing set* permet de s'assurer que la performance du modèle dans le premier jeu de données n'est pas le simple résultat d'un biais ou d'un artéfact statistique.

Les deux bases de données préliminaires ont ensuite été combinées (n = 3000) afin d'entraîner à nouveau le modèle sur une base de données plus large. Jumeler les deux bases de données pour l'entraînement du modèle final permet d'améliorer la qualité des prédictions: en effet, plus le nombre de répondants est élevé, meilleur est l'algorithme. Le modèle logistique multinomial a ensuite été implanté dans l'infrastructure de l'application Datagotchi, permettant alors la prédiction des intentions de vote des utilisateurs en fonction de l'équation de prédiction.

Après avoir complété le questionnaire de Datagotchi et après avoir obtenu ses prédictions, l'utilisateur valide (ou invalide) les prédictions d'intention de vote émises par l'application. De cette façon, l'algorithme d'apprentissage s'améliore constamment, sur la base des utilisateurs de Datagotchi.

Lien avec des bases de données massives

Enfin, la plateforme lie les données-utilisateurs sur les préférences musicales et sur les préférences de films[2] à des bases de données numériques massives. Plus précisément, les préférences musicales des utilisateurs sont liées à MusicBrainz, une encyclopédie musicale open source qui collecte des métadonnées sur les artistes musicaux et les met à la disposition du public. Grâce à un appel à l'interface de programmation applicative (API) de MusicBrainz, Datagotchi collecte des informations supplémentaires liées à l'artiste préféré(e) des utilisateurs (tels que le genre musical ou encore les années d'activité de l'artiste). Les préférences cinématographiques des utilisateurs sont quant à elles liées à Ombd, une base de données publique sur les films. Encore une fois, un appel à l'API de Ombd permet d'aller chercher des informations supplémentaires à partir du film préféré de l'utilisateur, telles que les artistes, ou encore l'année de réalisation. Ces méthodes novatrices de collectes de données permettent d'aller chercher rapidement des informations granulaires de façon rigoureuse et systématique. Ultimement, lorsque la quantité de données collectées permettra de faire des inférences robustes et valides, ces nouvelles informations seront intégrées aux analyses afin de tester leur effet prédictif dans les modèles. D'ici là, les variables liées aux préférences musicales et cinématographiques sont exclues des modèles de prédiction.

[2]Sur l'application, ces préférences sont collectées à travers les deux questions suivantes: Quel est votre groupe musical ou musicien(ne) préféré(e)? Choisissez votre album préféré! et Quel est votre film préféré?

Datagotchi est un projet ouvert et collaboratif. Pour toute question ou information, écrivez-nous à info@ datagotchi.com et il nous fera plaisir de discuter avec vous.

RÉFÉRENCES

Fréchet, N., Savoie, J., & Dufresne, Y. (2020). Analysis of text-analysis syllabi: Building a text-analysis syllabus using scaling. *PS: Political Science & Politics*, 53(2), 338-343.

Van Acker, Veronique. 2015. "Defining, Measuring, and Using the Lifestyle Concept in Modal Choice Research.'' *Transportation Resarch Board* 2495 (1): 74–82.

Wohlin, C. (2014). Guidelines for snowballing in systematic literature studies and a replication in software engineering. In Proceedings of the 18th international conference on evaluation and assessment in software engineering (pp.~1-10).

ANNEXE

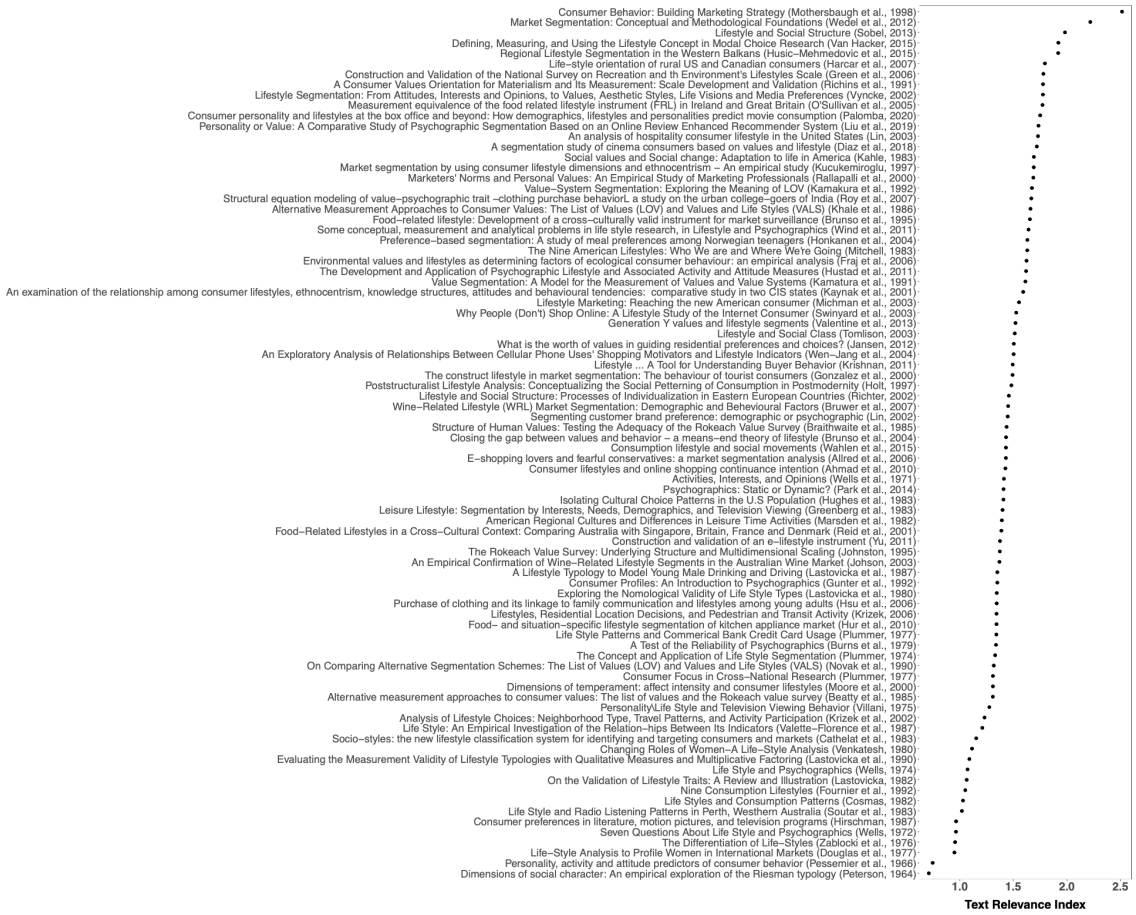

Figure 1: Liste des publications les plus pertinentes sur les habitudes de vie

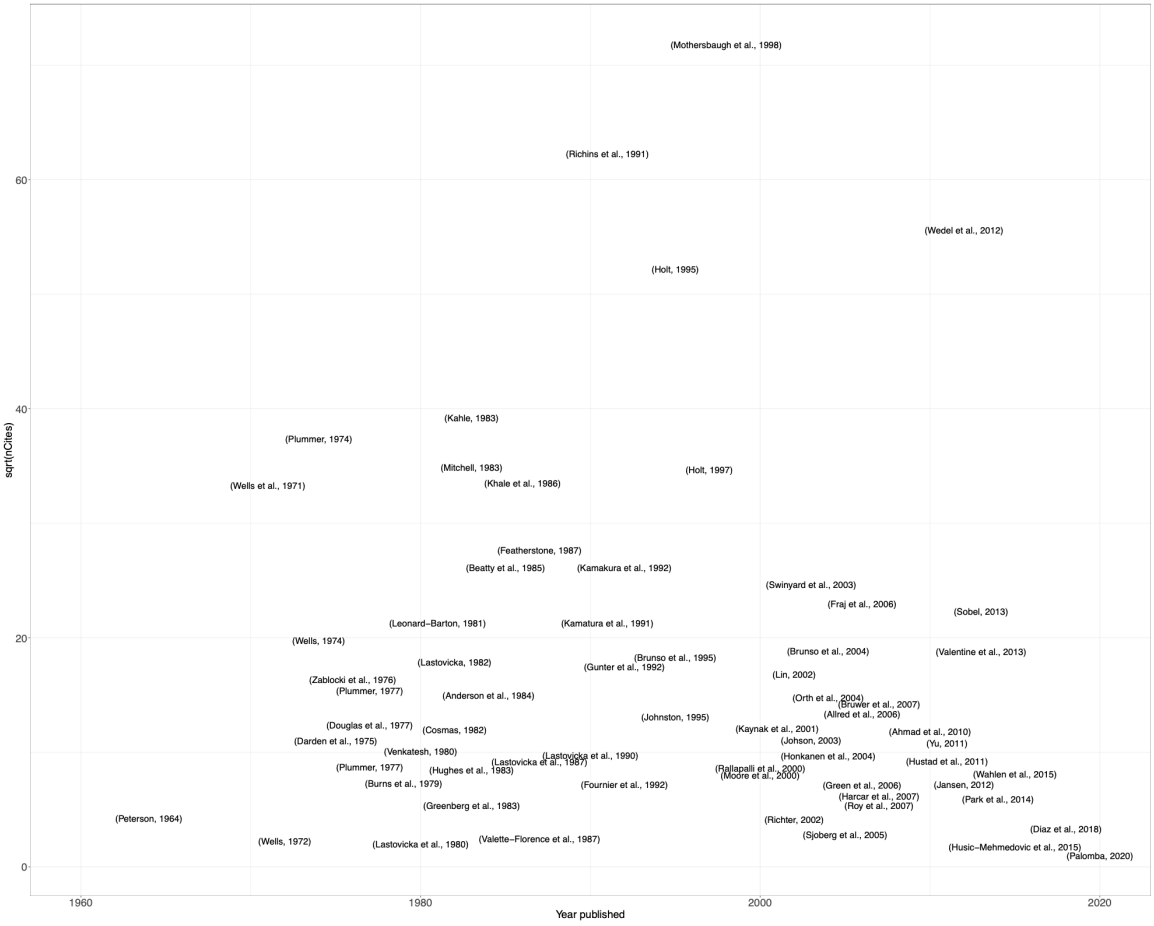

Figure 2: Les publications les plus pertinentes sur les habitudes de vie dans le temps