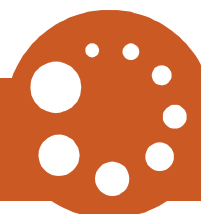




Plausi++: Contrôles automatiques de plausibilité pour vérifier la qualité et la fiabilité des données administratives et des données collectées par voie d'enquête

Sommaire de gestion

EXPERIMENTAL STATISTICS



Neuchâtel, 2024

Éditeur: Office fédéral de la statistique (OFS)
Renseignements: persfinhs@bfs.admin.ch
Rédaction: Mehmet Aksözen, OFS
Domaine: 15 éducation et science
Langue du texte original: Allemand
Traduction: section BILD-P

Concept de mise en page: section PUB
Téléchargement: www.statistique.ch
Copyright: OFS, Neuchâtel 2024
La reproduction est autorisée, sauf à des fins
commerciales, si la source est mentionnée

1 Situation initiale

Ce document se base sur le rapport "Plausi++ : Contrôles automatiques de plausibilité pour vérifier la qualité et la fiabilité des données administratives et des données collectées par voie d'enquête" [1] et doit servir de complément et de conclusion au projet pilote.

L'idée pour le projet pilote était d'utiliser un algorithme de machine learning pour compléter le contrôle de plausibilité jusqu'à présent manuel avec une partie automatisée et de trouver des erreurs potentielles non identifiées jusqu'à présent (cas problématiques).

Les données utilisées se basent sur les données des statistiques du personnel des hautes écoles universitaires, complétées par des données issues des statistiques des étudiants et des examens, ainsi que par des chiffres clés mathématiques (voir [1], p.6).

Le modèle renvoie une catégorie de personnel prédite pour chaque ensemble de données. Si cette catégorie de personnel prédite ne correspondait pas à celle fournie (cas problématique), nous avons cherché des candidats de variables dont les valeurs pourraient avoir contribué à cette différence.

2 Objectifs

Dans le rapport précédent [1], p. 6, plusieurs points ont été énumérés pour la suite du projet. Ceux-ci correspondent aux objectifs des travaux présentés ci-dessous :

1. Test du mécanisme de feedback : le mécanisme de feedback créé dans une première ébauche donne une indication sur les cas problématiques (candidats de variables qui peuvent être erronés), qui seront ensuite vérifiés par les livreurs de données.
2. L'implication des livreurs de données : Les livreurs de données reçoivent le feedback sur les erreurs potentielles et les vérifient. Les livreurs de données expliquent pourquoi les données sont correctes ou livrent des données corrigées. En outre, les différences organisationnelles entre les livreurs de données peuvent se refléter dans les données par des écarts systématiques qui peuvent à leur tour donner lieu à des groupes de cas problématiques.
3. Tenir compte des changements dans les hautes écoles universitaires : Outre les cas problématiques, les révisions ou autres changements au sein des hautes écoles universitaires peuvent entraîner une modification des structures de données (par exemple, choix d'une catégorie de personnel différente de celle utilisée jusqu'à présent en raison d'adaptations internes), de sorte que le modèle peut trouver davantage de cas problématiques pouvant être attribués à ces changements.
4. Suivi du modèle : si des différences dans les données sont mesurées entre deux années pendant l'enquête, l'algorithme est entraîné à nouveau avec les données disponibles. Sinon, un nouvel entraînement avec l'algorithme a lieu après la fin de l'enquête, qui est comparée à l'ancien modèle.

3 Réalisation des objectifs

Objectif 1 : Quatre groupes de cas problématiques potentiels ont été envoyés aux livreurs de données, ce qui représente environ 28% de tous les cas. Afin d'alléger la charge de travail des livreurs de données, le mécanisme de feedback a été vérifié avant l'envoi. Ainsi, des cas problématiques pour lesquels la probabilité de trouver une erreur semblait particulièrement élevée ont été signalés. De plus, des informations supplémentaires ont été rajoutées à ces cas.

Objectif 2 : Les livreurs de données ont pu confirmer que tous les cas problématiques sélectionnés (erreurs potentielles) étaient corrects, même ceux qui se produisent parfois rarement. D'autres retours d'information de la part des livreurs de données sur un autre groupe de cas potentiellement problématiques (32% de tous les cas sont liés à l'administration des hautes écoles) ont permis de constater qu'il peut y avoir plusieurs raisons structurelles aux écarts.

- La collecte des données est différente (calcul ou enquête complète).
- L'administration des institutions est différente, ce qui a un impact différent sur les catégories de personnel.
- L'offre de prestations des livreurs de données a un focus différent (plus d'enseignement) par rapport à la majorité (plus de recherche et développement).

Objectif 3 : Avec le calcul d'un indice de stabilité de la population, un module a été ajouté afin de vérifier la répartition entre les années et d'évaluer la pertinence de l'utilisation de l'algorithme précédent. La répartition des données par catégorie de personnel, par haute école et pour les autres données fournies ne différait pas sensiblement d'une année à l'autre.

Objectif 4 : Après chaque relevé, l'algorithme a été réentraîné avec les données les plus récentes. La précision très élevée de ces modèles réentraînés chaque année est restée stable. La précision des modèles de prédiction de la catégorie de personnel a également peu changé lorsque les modèles ont été appliqués aux données d'autres années.

La qualité des données de la statistique du personnel est très élevée. Les différences entre les hautes écoles sont de nature technique et structurelle ; elles ont été discutées avec les hautes écoles à la fin du projet. En cas de besoin de la part des hautes écoles, l'algorithme peut être réutilisé.

4 Organisation du projet

La poursuite du projet a coïncidé avec une phase où les livreurs de données et la gestion de projet avaient peu de ressources pour effectuer un test (surtout lors de la pandémie de COVID-19). Afin de réduire l'effort des institutions, la gestion de projet a investi un peu plus de temps et de ressources et a rassemblé une sélection de cas problématiques potentiels ayant une forte probabilité de devoir être corrigés. En raison de cette situation un peu plus difficile à planifier, la gestion de projet a réduit au minimum l'implication d'autres collaborateurs et collaboratrices des sections processus de formation ou système de formation. La préparation des données a été effectuée par le groupe IT de la section processus de formation.

5 Références

[1] Ruiz, C. et al., 2019: «Plausi++: Automatische Plausibilitätsprüfung der Qualität und Zuverlässigkeit von Administrativ- und Umfragedaten»,
site web: <https://www.bfs.admin.ch/bfs/de/home.assetdetail.9847917.html> (dernier accès 25.03.2024)

[2] Yurdakul, B., 2018: «Statistical Properties of Population Stability Index», Dissertation,
site web: <https://scholarworks.wmich.edu/cgi/viewcontent.cgi?article=4249&context=dissertations> (dernier accès 25.03.2024)