Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement des Innern EDI
Bundesamt für Statistik BFS

Plausi++: Automatische Plausibilitätsprüfung der Qualität und Zuverlässigkeit von Administrativ- und Umfragedaten

Management Summary

EXPERIMENTAL STATISTICS

Neuchâtel, 2024

Herausgeber:	Bundesamt für Statistik (BFS)	**Layoutkonzept:**	Sektion PUB
Auskunft:	persfinhs@bfs.admin.ch	**Download:**	www.statistik.ch
Redaktion:	Mehmet Aksözen, BFS	**Copyright:**	BFS, Neuchâtel 2024
Themenbereich:	15 Bildung und Wissenschaft		Wiedergabe unter Angabe der Quelle für nichtkommerzielle Nutzung gestattet
Originaltext:	Deutsch		
Übersetzung:	Sektion BILD-P		

1 Ausgangslage

Dieses Dokument baut auf dem Bericht «Plausi++: Automatische Plausibilitätsprüfung der Qualität und Zuverlässigkeit von Administrativ- und Umfragedaten» [1] auf und soll als Ergänzung und Abschluss des Pilotprojekts dienen.

Die Idee für das Pilotprojekts war es, durch einen Machine Learning Algorithmus die bislang manuelle Plausibilisierung mit einem automatisierten Teil zu ergänzen und potenzielle bislang nicht identifizierte Fehler (Problemfälle) zu finden.

Die genutzten Daten basieren auf den Daten der Personalstatistik der universitären Hochschulen, ergänzt um Daten aus den Statistiken der Studierenden und Examen sowie mathematischen Kennzahlen (siehe [1], S.6).

Das Modell gibt für jeden Datensatz eine vorhergesagte Personalkategorie zurück. Wenn diese vorhergesagte Personalkategorie nicht der gelieferten entsprach (Problemfall), wurde nach Kandidaten von Variablen gesucht, deren Werte zu diesem Unterschied beigetragen haben könnten.

2 Zielsetzung

Im Ausblick des vorherigen Berichts [1], S. 6 werden mehrere Punkte für die Weiterführung des Projekts aufgelistet. Diese entsprechen der Zielsetzung der hier vorgestellten Arbeiten und betreffen wie folgt:

1. Test des Feedback-Mechanismus: Der in einem ersten Entwurf erstellte Feedback-Mechanismus gibt einen Hinweis zu Problemfällen (Kandidaten von Variablen, die fehlerhaft sein können), welche dann von den Datenlieferanten überprüft werden.
2. Einbezug von Datenlieferanten: Die Datenlieferanten erhalten das Feedback zu möglichen Kandidaten und überprüfen diese. Die Datenlieferanten erläutern, weshalb die Daten korrekt sind, oder liefern korrigierte Daten. Zudem können sich organisatorische Unterschiede zwischen den Datenlieferanten in den Daten durch systematische Abweichungen abbilden, die wiederum Gruppen von Problemfällen ergeben können.
3. Change Management an den universitären Hochschulen berücksichtigen: Neben den Problemfällen können Revisionen oder andere Änderungen an den Hochschulen dazu führen, dass sich die Datenstrukturen ändern (z.B. Wahl einer anderen Personalkategorie als bislang wegen interner Anpassungen), so dass das Modell vermehrt Problemfälle finden kann, die auf diese Änderungen zurückführbar sind.
4. Monitoring des Modells: Wenn Unterschiede in den Daten zwischen zwei Jahren während der Erhebung gemessen werden, wird der Algorithmus mit den verfügbaren Daten neu trainiert. Ansonsten erfolgt ein neues Training mit dem Algorithmus nach dem Abschluss der Erhebung, welcher mit dem alten Modell verglichen wird.

3 Zielerreichung

Ziel 1: Den Datenlieferanten wurden vier Gruppen von möglichen Problemfällen zugestellt, die etwa 28% aller Fälle ausmachen. Um den Aufwand bei den Datenlieferanten möglichst gering zu halten wurde der Feedback-Mechanismus zusätzlich vor dem Versand überprüft. So wurden Problemfälle gemeldet bei denen die Wahrscheinlichkeit, einen Fehler zu finden, besonders hoch erschien. Zusätzlich wurden zu diesen Fällen weitere Angaben hinzugezogen.

Ziel 2: Die Datenlieferanten konnten alle ausgewählten Problemfälle (potenzielle Fehler) als korrekt zurückmelden, auch solche die zum Teil selten vorkommen. Aus weiteren Rückmeldungen der Datenlieferanten zu einer weiteren Gruppe von möglichen Problemfällen (32% aller Fälle mit Bezug zur Verwaltung der Hochschule) stellte sich heraus, dass es mehrere strukturelle Gründe für Abweichungen geben kann:

- Die Erfassung der Daten erfolgt unterschiedlich (Berechnung oder Vollerhebung).
- Die Verwaltung der Institutionen ist unterschiedlich gestaltet, was sich auf die Personalkategorien unterschiedlich auswirkt.
- Das Leistungsangebot der Datenlieferanten hat einen unterschiedlichen Fokus (mehr Lehre) gegenüber der Mehrheit (mehr Forschung und Entwicklung).

Ziel 3: Mit der Berechnung eines Population Stability Index [2] wurde ein Modul ergänzt um die Verteilung zwischen den Jahren zu überprüfen, um die Einsatzfähigkeit des vorherigen Algorithmus zu beurteilen. Die Verteilung der Daten pro Personalkategorie, pro Hochschule und bei weiteren gelieferten Daten unterschied sich zwischen den Jahren nicht merklich.

Ziel 4: Nach jeder Erhebung wurde der Algorithmus mit den aktuellsten Daten neu trainiert. Die sehr hohe Genauigkeit dieser jährlich neu trainierten Modelle blieb stabil. Die Genauigkeit der Modelle bei der Vorhersage der Personalkategorie änderte sich auch kaum, wenn Modelle auf Daten anderer Jahre angewandt wurden.

Die Datenqualität der Personalstatistik ist sehr hoch. Die Unterschiede an den Hochschulen sind erhebungstechnischer und struktureller Natur und wurden zum Abschluss des Projekts mit den Hochschulen besprochen. Bei Bedarf seitens der Hochschulen kann der Algorithmus wieder genutzt werden.

4 Projektorganisation

Die Weiterführung des Projekts fiel in eine Phase, in der sowohl die Datenlieferanten als auch die Projektleitung wenig Ressourcen für einen Testdurchlauf hatten (v.a. die Corona-Pandemie). Um den Aufwand für die Institutionen auf ein geringes Mass zu reduzieren, hat die Projektleitung etwas mehr Zeit und Ressourcen investiert und eine Auswahl an potenziellen Problemfällen mit hoher Wahrscheinlichkeit für eine Korrektur zusammengestellt. Auf Grund dieser etwas schwieriger zu planenden Situation hat die Projektleitung den Einbezug weiterer Mitarbeitenden aus den Sektionen BILD-P oder BILD-S auf die notwendige Datenaufbereitung seitens der IT der Sektion BILD-P auf das Minimale reduziert.

5 Referenzen

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

[2] Yurdakul, B., 2018: «Statistical Properties of Population Stability Index», Dissertation,
Webseite: https://scholarworks.wmich.edu/cgi/viewcontent.cgi?article=4249&context=dissertations
(letzter Zugriff 25.03.2024)