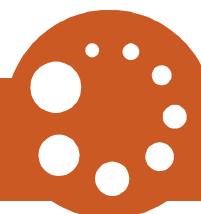




Das Dateninnovationsprojekt «NOGAuto»

Management Summary

EXPERIMENTAL STATISTICS



Neuchâtel, 2021

Herausgeber: Bundesamt für Statistik (BFS)
Auskunft: cindia.ducsfez@bfs.admin.ch, tel. +41 58 463 63 37
Redaktion: Lorenz Helbling (lorenz.helbling@bfs.admin.ch)
Themenbereich: 00 Statistische Grundlagen
Originaltext: Englisch
Übersetzung: Sektion URD

Layoutkonzept: Sektion DIAM
Download: www.experimental.bfs.admin.ch
Copyright: BFS, Neuchâtel 2021
Wiedergabe unter Angabe der Quelle
für nichtkommerzielle Nutzung gestattet

Das Dateninnovationsprojekt "NOGAuto"

Management Summary

Dieses Dokument beschreibt, wie das maschinelle Lernen Projekt "NOGAuto" von einem Hochschulpraktikanten im Rahmen seiner Masterarbeit an der Universität von Genf - Geneva School of Economics and Management (GSEM) Fakultät im Master of Science in Business Analytics während seines zehnmonatigen Praktikums beim Bundesamt für Statistik (BFS) durchgeführt wurde.

In 2017 lancierte das BFS seine Dateninnovationsstrategie, um den technologischen Entwicklungen voraus zu sein. Das Projekt "NOGAuto", das von der Sektion Unternehmensregisterdaten (URD) eingereicht wurde und eines der fünf wichtigsten Projekte war, wurde im Einklang mit der Dateninnovationsstrategie des BFS beibehalten, mit dem Ziel, die bestehende Basisproduktion der amtlichen Statistik des Amtes zu erweitern und/oder zu ergänzen.

Als wesentliches Element der statistischen Produktion der Allgemeinen Systematik der Wirtschaftszweige (NOGA) des BFS wurde das Projekt "NOGAuto" durchgeführt, um die Kodierung der wirtschaftlichen Tätigkeit der Unternehmen zu automatisieren. Dieses Vorhersagesystem zielt darauf ab, die Zuordnung oder, genauer gesagt, die Vorhersage von Wirtschaftszweigcodes von Unternehmen zu automatisieren, und zwar zunächst auf der Grundlage von Informationen, die im BFS bereits verfügbar sind (z. B. Erhebungsdaten, Handelsregisterbeschreibungen, Erläuterungen zur Klassifizierung), und unter Verwendung von Methoden des "supervised machine learning" im Interesse der Entlastung der Unternehmen und zur Unterstützung der Kodierung.

Durch das Projekt "NOGAuto" könnten alle Statistiknutzer des BFS von:

- einem "Tool", das automatisch die NOGA-Codes der Unternehmen überprüft;
- einer Standardisierung der Kodierung;
- einer Minimierung des Interpretationsfaktors;
- einer Optimierung des Zeitaufwands für die Kodifizierung;
- einer Verbesserung der Qualität der NOGA-Kodierung und somit der gesamten Unternehmensstatistik profitieren.

Im ersten Teil des Dokuments unterstreicht die Masterarbeit die Bedeutung der Anwendung eines klar definierten Prozesses, bei dem das Produkt drei Phasen durchläuft (Prototype, Minimal Viable Product und Production), und eines angemessenen Verständnisses der Problematik. In einem zweiten Teil werden die Techniken des Data Mining (z. B. text cleaning usw.) und des Feature Engineering (z. B. word embedding usw.) beschrieben, die zur Vorbereitung der Daten für den Modellierungsteil verwendet wurden, in dem Methoden des machine learning (z. B. gradient boosting machine) eingesetzt wurden. Schliesslich werden im Kapitel "lessons learnt" einige der Höhepunkte des Projekts diskutiert, wie die Präsentation an der europäischen Konferenz New Techniques & Technologies for Statistics (NTTS) 2021, und welche Verbesserungen für die Zukunft des Projekts vorgenommen werden können.