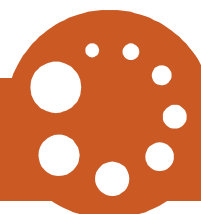




Le projet d'innovation des données «NOGAuto»

Management Summary

EXPERIMENTAL STATISTICS



Neuchâtel, 2021

Éditeur:	Office fédéral de la statistique (OFS)	Concept de mise en page:	Section DIAM
Renseignements:	cindia.ducsfez@bfs.admin.ch, tel. +41 58 463 63 37	Téléchargement:	www.statistique.ch
Rédaction:	Lorenz Helbling (lorenz.helbling@bfs.admin.ch)	Copyright:	OFS, Neuchâtel 2021
Domaine:	00 Bases statistiques des généralités		La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée
Langue du texte original:	Anglais		
Traduction:	Section URD		

Le projet d'innovation des données "NOGAuto"

Management summary

Ce document décrit comment le projet d'apprentissage automatique "NOGAuto" a été mené pendant les dix mois de stage à l'Office fédéral de la statistique (OFS) par un stagiaire haute école dans le cadre de son travail de Master en Master of Science in Business Analytics à la faculté Geneva School of Economics and Management (GSEM) de l'Université de Genève.

En 2017, l'OFS a lancé sa stratégie d'innovation en matière de données pour rester à la pointe de l'évolution technologique. Le projet, qui a été soumis par la section Données des registres des entreprises (URD) et qui était l'un des cinq principaux projets, a été retenu conformément à la stratégie d'innovation en matière de données de l'OFS dans le but d'augmenter et/ou de compléter la production statistique officielle de base existante à l'office.

Selon l'élément essentiel de la production statistique de la Nomenclature Générale des Activités Economiques (NOGA) à l'OFS, le projet "NOGAuto" a été réalisé pour automatiser le codage de l'activité économique des entreprises. Ce système de prédiction vise à automatiser l'attribution ou, plus précisément, la prédiction des codes d'activité économique aux entreprises, en se basant dans un premier temps sur les informations déjà disponibles au sein de l'OFS (par exemple, les données d'enquête, les descriptions du registre du commerce, les notes explicatives des classifications) et en utilisant des méthodes d'"apprentissage automatique supervisé" dans l'intérêt de décharger les entreprises et pour soutenir le codage.

Grâce au projet "NOGAuto", tous les utilisateurs de statistiques de l'OFS pourraient bénéficier:

- un "outil" qui vérifie automatiquement les codes NOGA des entreprises;
- une standardisation de la codification;
- une minimisation du facteur d'interprétation;
- une optimisation du temps consacré à la codification;
- une amélioration de la qualité de la codification NOGA et, par conséquent, de toutes les statistiques sur les entreprises.

Dans la première partie du document, le mémoire de Master souligne l'importance de l'application d'un processus bien défini, où le produit passe par trois phases (Prototype, Minimal Viable Product et Production) et une bonne compréhension du contexte de la problématique. Dans une deuxième partie, le document décrit les techniques de data mining (ex. text cleaning, etc.) ainsi que de feature engineering (ex. word embedding, etc.) qui ont été utilisées pour préparer les données pour la partie modélisation, où des méthodes de machine learning (ex. gradient boosting machine) ont été appliquées. Enfin, dans le chapitre "lessons learnt", certains des points culminants du projet seront abordés, comme sa présentation à la conférence européenne New Techniques & Technologies for Statistics (NTTS) 2021, ainsi que ce qui peut être amélioré pour l'avenir du projet.