## Technical overview

The project consists of four parts: Scraping data from websites, extracting useful features from the scraped data, training a ML model to predict the label for a website, and creating a command line interface and a web user interface so that people who are less comfortable with command line interface, such as data analysts and data scientists, are able to use the platform.

Whole project is done locally, without databases or external services.

The scraping is done for a list of webshop URL addresses provided by F-secure. The URL addresses have been already classified as fake or legitimate. For the scraping, we use, among other tools, the ScaPy library for creating and sending packets, and the playwright library for rendered scraping. Additionally, for efficiency, asynchronous code or threads are needed. To ensure safety during scraping and testing virtual machines are used, as well as VPN.

The feature extraction includes some of the features from a master thesis “Detection of fake web-shops in the .be zone using active learning” by Jules Dejaeghere, as well as someadditional features. Machine learning model is a Random Forest Classifier. Model has features based on html extraction information, as well as domain and SSL certificate extraction. Model makes a prediction and provides the classification result to the UI. Accuracy of the model is 0,86.

The program has two interfaces, a command line interface and a graphical user interface.

The graphical user interface is run only locally. In web UI users can input URL addresses, as a file or in a textfield. It is possible to change some of the options of scraping and extraction.Classification result, some additional data and screenshots are shown in the UI.

## 