Implemented feature: Web scraping job data and store in database

As part of the Pathfinder project, the first major feature implemented was a web scraping component designed to extract job information from the job search websites. Apart from the Kaggle datasets this feature is an important part of the data collection process for the Pathfinder career recommendation system.

In the original proposal, the database was planned to be MySQL or H2. However, due to library dependency issues and setup difficulties, We decided to use SQLite instead. SQLite was easier to configure, lightweight, and worked well with Python. This change allowed faster development and testing without affecting the functionality of the project.

The main objective of this feature is collecting real job posting data such as job title, company name, location, and description and store the collected data in a structured database to prepare data that can later be used for recommendation and skill analysis.

We used selenium, webdriver\_manager libraries and implemented the scraper to handles cookie consent pop-ups, random delays to mimic human browsing, and automatically save the extracted data into the database.

We assume the indeed website maintains a consistent structure for job postings and use ethical scraping principles respecting delays and not overloading the site. We tried different scraping methods using Selenium and BeautifulSoup and selected Selenium for its ability to handle dynamic JavaScript-rendered content.

A screenshot of a computer code

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.