International Internship Report

Development of a CSV Validation and Correction Tool for Process Mining

Author:

BOUKHRIS Mohamed Lyazid

Academic tutor:

Dr. Gema Ibañez Sanchez

Host institution:

ITACA Institute, Universitat Politècnica de València

SABIEN Research Group

Internship period:

19 May 2025 – 8 August 2025

Contents

Chapter 1

Host Institution and Research Group

1.1 ITACA Institute

The ITACA Institute (Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones Avanzadas) is a research centre specialised in information and communication technologies.

Its mission is to put scientific innovation at the service of society, with a strong focus on health, well-being, communication systems and transport. The institute gathers more than one hundred researchers and participates in numerous national and European projects in areas such as biomedical engineering, communication systems, electromagnetic compatibility, navigation and transport.

1.2 SABIEN Research Group

My internship took place within the SABIEN research group at ITACA. SABIEN focuses on applying information technologies to health and well-being. The team develops digital tools and innovative solutions to support healthcare professionals and improve the quality of care.

The group is particularly active in:

- Data analysis and **process mining** for healthcare processes,
- Decision support systems for clinicians and managers,
- Collaborative projects with hospitals and European partners.

Within this context, my mission was directly linked to the preparation and improvement of data used for process mining in a healthcare setting.

Chapter 2

Project Overview

2.1 Context and Motivation

Process mining techniques rely heavily on the quality of event data extracted from information systems. In practice, these data are often stored or exported as CSV or extended CSV (eCSV) files.

However, raw CSV files typically suffer from:

- inconsistent delimiters,
- incorrect column types,
- noisy or inconsistent categorical values,
- missing fields or structural errors.

These issues make process discovery and conformance checking more difficult, and they increase the time needed to prepare datasets before feeding them into process mining tools such as *PMapp* (Process Mining application) used by the team.

2.2 Mission Objective

The main mission of my internship was to design and develop a desktop tool capable of:

- validating the **structure** of CSV and eCSV files,
- detecting common errors in delimiter usage and column types,
- identifying frequent or inconsistent values,
- guiding the user to correct these problems,
- exporting a cleaned CSV file and a detailed validation report.

The target users are data analysts and researchers who prepare event logs for process mining. The tool aims to reduce manual work, standardise data preparation and improve the overall quality of logs used by PMapp.

2.3 Technologies and Methods

To implement the CSV Structural Validator application, I used the following technologies:

- **C# and WPF** for the desktop application and graphical user interface.
- **.NET** for the internal logic, file parsing and validation pipeline.
- **Embedding-based algorithms** to automatically recognise and match columns based on their semantic similarity.
- A **type detection engine** combining rule-based heuristics and basic machine-learning techniques to infer data types (string, date, integer, etc.).
- **Dictionaries and embeddings** to normalise categorical values, especially when multiple spellings or variants represent the same concept.

The result is an end-to-end validation pipeline where the user remains in control: the system proposes suggestions, and the user validates or adjusts them.

Chapter 3

Application Workflow

The CSV Structural Validator application is organised as a sequence of clear steps, each guided by the graphical interface. This section describes the workflow followed by the user, illustrated with screenshots from the final application.

3.1 Step 1 – Starting the Validation and Selecting a File

The main window welcomes the user and invites them to start the validation by clicking on *Start CSV Validation*. A file selection dialog then appears, allowing the user to choose a CSV or eCSV file from the file system.

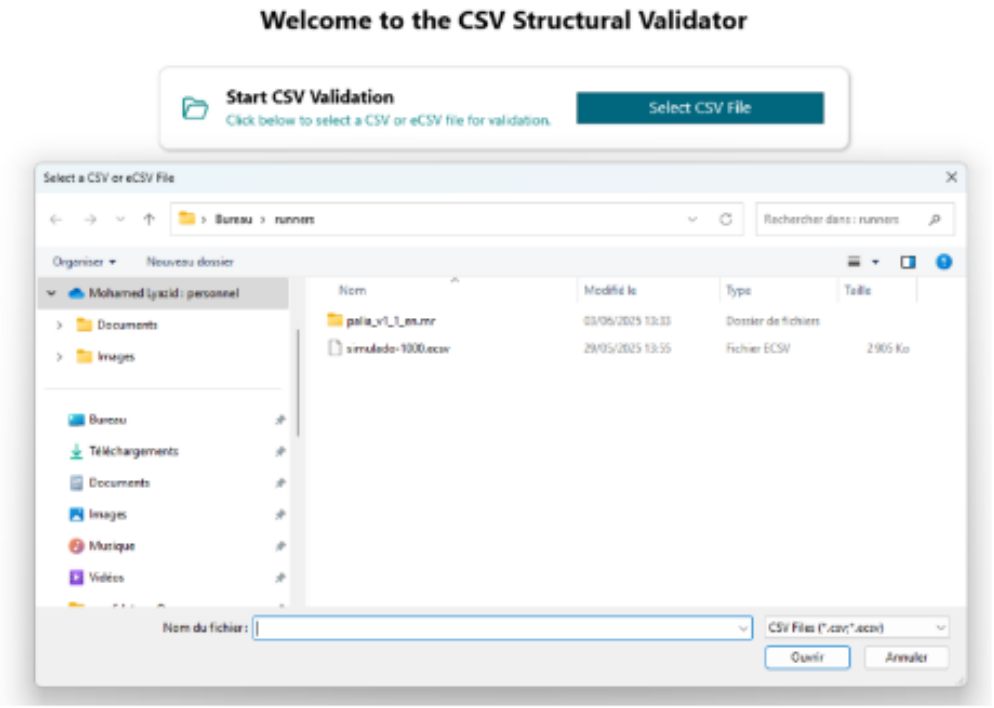

Figure 3.1: Step 1 – Starting CSV validation and selecting the CSV file.

Once the file is selected, the main interface confirms that the CSV file has been loaded and enables the next step, which is delimiter validation.

3.2 Step 2 – Delimiter Validation

The second step focuses on verifying the column delimiter used in the file. A dedicated window opens and displays a preview of the first lines of the CSV file, with the current delimiter applied.

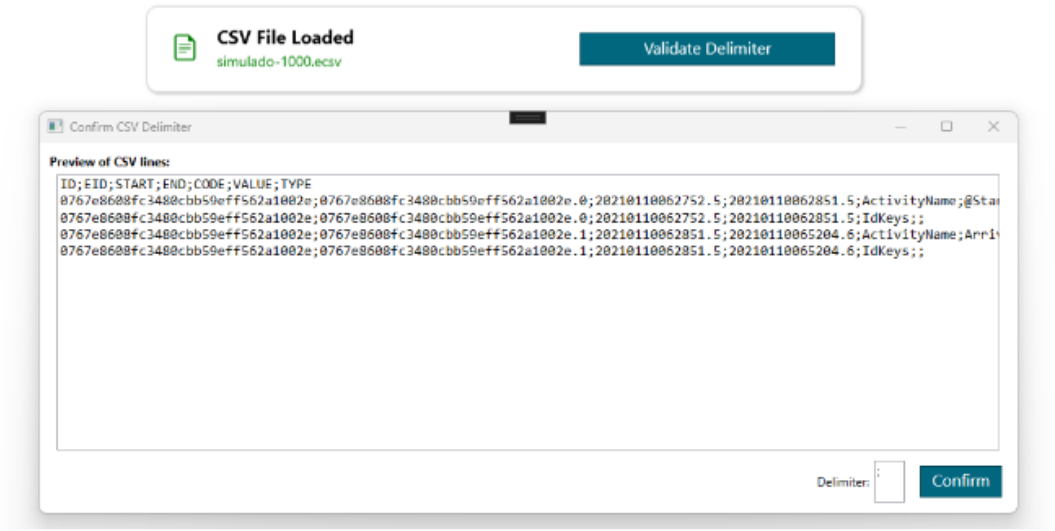

Figure 3.2: Step 2 – Confirming and validating the CSV delimiter.

The user can:

- visually check that columns are correctly separated,
- modify the delimiter manually if the automatic detection is wrong,
- confirm the delimiter to proceed.

This step ensures that the subsequent analysis is based on a correct interpretation of the file structure.

3.3 Step 3 – Column Type Detection and Validation

After the delimiter is confirmed, the application analyses the content of each column and proposes a data type (e.g. string, date, integer). The user is presented with a table containing:

- the preview of the first rows of each column,
- a drop-down selector for the inferred type.

The user can adjust the suggested types if necessary. This manual confirmation is crucial to guarantee that:

- timestamps are correctly interpreted as dates,
- identifiers remain strings,
- numerical fields are consistently treated as such.

This step directly contributes to the quality of subsequent process mining analyses.

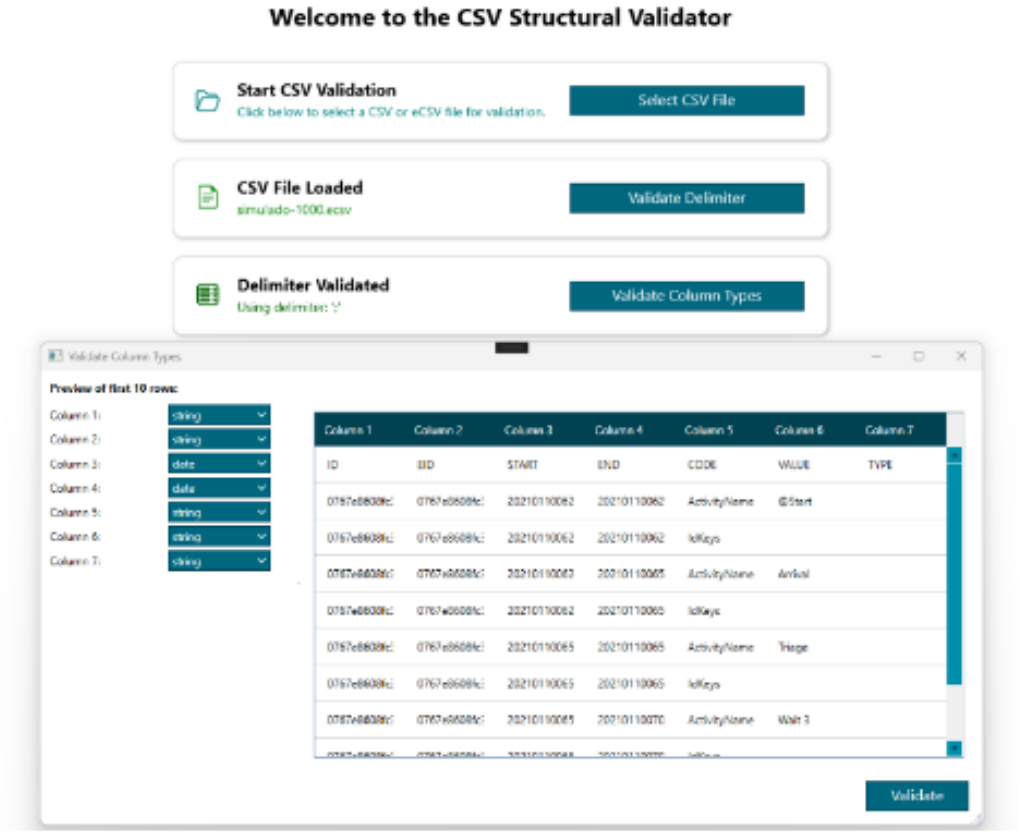

Figure 3.3: Step 3 – Validating inferred column types.

3.4 Step 4 – Detection of Frequent and Inconsistent Values

The fourth step aims at identifying recurring or potentially problematic values in each column. The tool scans the data and computes frequency distributions.

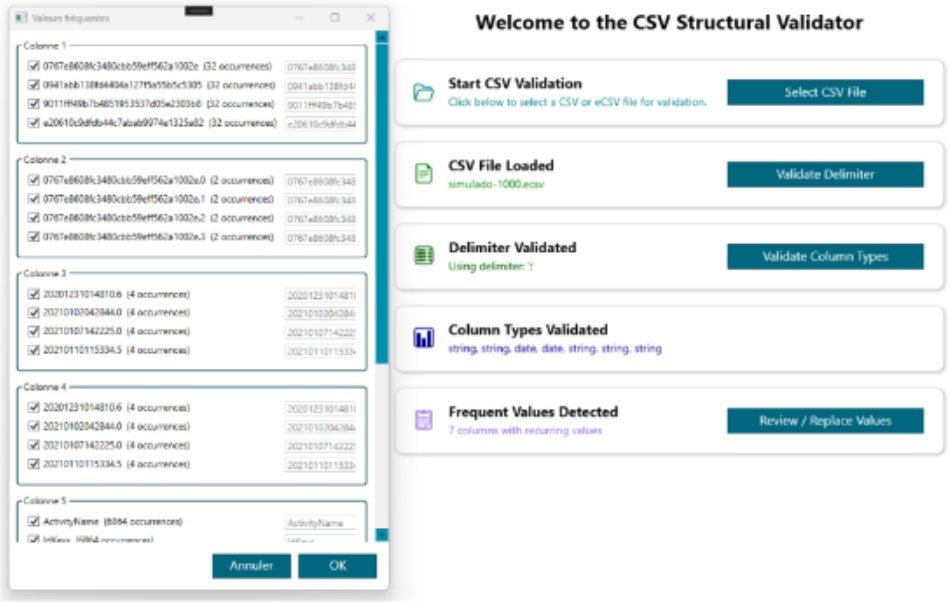

Figure 3.4: Step 4 – Detecting frequent values and preparing normalisation.

The interface displays, for each column:

- the most frequent values with their number of occurrences,
- candidate values that might be merged (e.g. different spellings of the same activity name),
- controls to select values for replacement or normalisation.

This allows the user to:

- harmonise categorical labels (e.g. *Arrival* vs. *Arrive*),
- detect obvious anomalies,
- decide how to clean the dataset while keeping full control over the changes.

3.5 Step 5 – Validation Summary and Export

Once the previous steps are completed, the application presents a global summary of the validation process. The main window displays:

- a confirmation that the column types have been validated,
- the number of columns with frequent values detected,

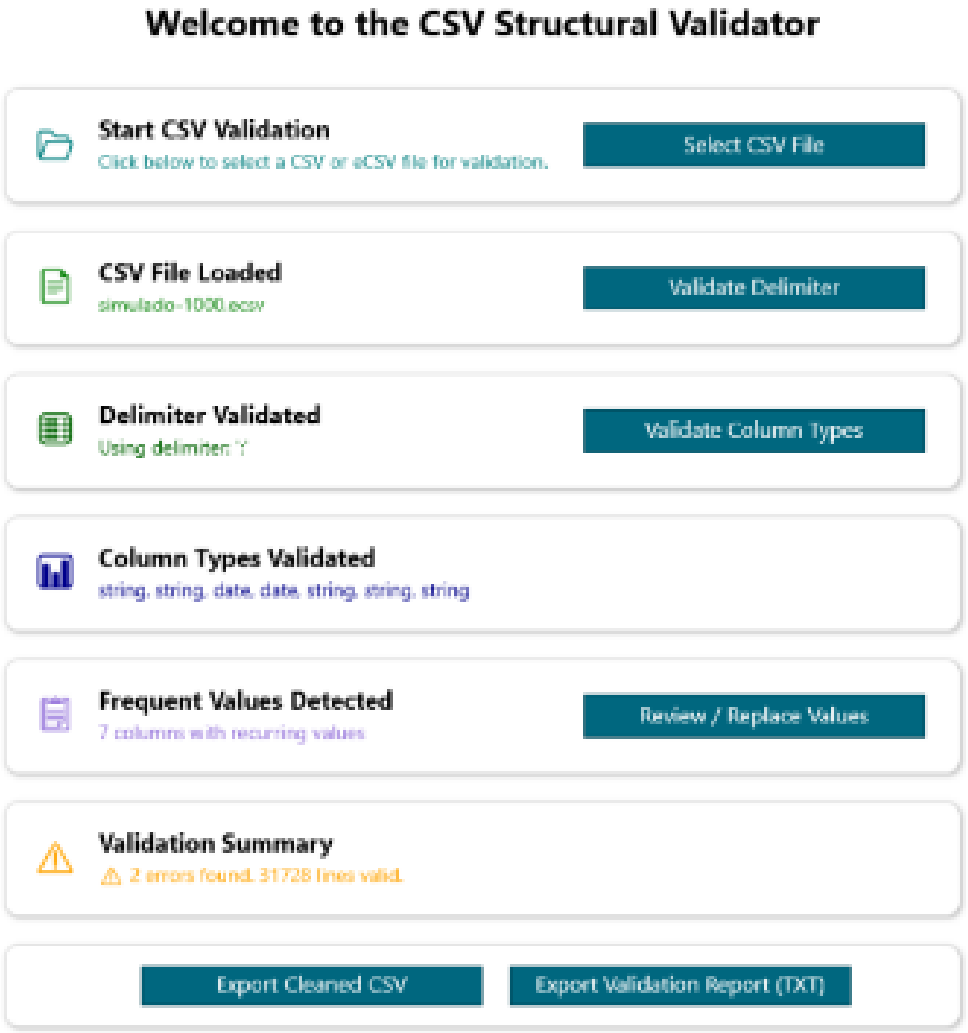

Figure 3.5: Step 5 – Validation summary and export of results.

- a validation summary including the number of valid lines, detected errors and corrections applied.

At this stage, the user can:

- export a **cleaned CSV** file, ready to be imported into PMapp,
- export a **textual validation report** (TXT) containing the details of the detected issues and applied corrections.

This final step completes the validation pipeline and provides full transparency on the cleaning operations performed.

Chapter 4

Sustainable Development and CSR at ITACA

Beyond the technical work, the internship also offered insight into how the host institution addresses sustainability and corporate social responsibility.

4.1 Environmental Considerations

The ITACA Institute is located on the campus of the Universitat Politècnica de València, in modern buildings designed to reduce energy consumption. The campus includes green areas and encourages environmentally friendly practices.

While ITACA's research activity does not involve heavy industrial processes, it uses a significant amount of computing equipment. To mitigate the environmental impact, the institute:

- optimises the use of computing resources,
- shares hardware between research groups whenever possible,
- promotes energy-efficient practices in offices and laboratories.

4.2 Social and Inclusive Dimension

On the social side, ITACA follows the inclusive policy of the university. The institute:

- facilitates the integration of people with disabilities,
- promotes cultural diversity by hosting researchers and students from different countries,
- encourages collaboration and knowledge sharing within international teams.

4.3 Contribution to Societal Challenges

Many ITACA projects aim to address concrete societal needs. Examples include:

- digital health solutions to improve the work of healthcare professionals and the quality of patient care,

- tools for better management of resources,
- research on the role of technologies in tackling challenges such as climate change and population ageing.

My project contributes indirectly to these goals by improving the quality of data used in health-related process mining studies.

Chapter 5

Skills and Personal Development

5.1 Professional Skills

This internship significantly strengthened my technical and methodological skills:

- **Software development:** I improved my proficiency in C#, WPF and .NET for building desktop applications.
- **Data processing pipeline:** I learned how to design and implement a complete data-processing chain, from file parsing to validation and export.
- **Applied AI techniques:** I used embeddings, type detection and value normalisation techniques in a concrete project.
- **End-to-end project management:** I followed all stages of the project, from requirements analysis to delivery of a functional solution, including documentation and iterative feedback.
- **Data quality in process mining:** I gained a deeper understanding of how data quality impacts process mining and, more generally, data-driven projects in digital health.

5.2 Personal Skills

On a personal level, this international experience in Valencia helped me grow in several ways:

- **Adaptation to a new environment:** I discovered a different cultural and professional context and adapted to working in an international research team.
- **Communication:** I improved my communication skills in English and Spanish, both in informal discussions and in technical meetings.
- **Autonomy and responsibility:** Managing my own tasks and presenting my progress every week to my supervisor and the team increased my sense of responsibility and confidence.
- **Career orientation:** This internship confirmed my motivation to work in cybersecurity and in the development of data-related tools.

Chapter 6

Conclusion

During this internship at the ITACA Institute, I designed and implemented the CSV Structural Validator application, a desktop tool dedicated to the validation and correction of CSV and eCSV files for process mining.

The project resulted in:

- a complete validation pipeline (delimiter, column types, frequent values, summary and export),
- an interactive WPF interface keeping the user at the centre of the process,
- a combination of rule-based and AI techniques for better data quality,
- documentation and a working prototype that can be integrated into the PMapp ecosystem.

Beyond the technical achievements, this mission gave me the opportunity to work in a dynamic research environment, to contribute to projects with a real impact on digital health, and to develop both professional and personal skills.

Overall, this experience has been extremely valuable and has strengthened my interest in data engineering, process mining and cybersecurity.