



Automatic Classification and Coding for the National Employment Survey (ENE)

Department of Labour Statistics
Technical Subdirectorate
National Statistics Institute

21st International Conference of Labour Statisticians

1. Problem and solution
2. Methodological background
3. Automatic Classification Procedure for ENE Texts
4. Quality control
5. Institutional API
6. Next steps

1. Problem and solution

Original requirement → Manual coding → Achieve greater efficiency

COLLECTION

CODING

DATA ENTRY

CONSTRUCTION OF THE DATABASE FOR
TRAINING THE MODEL



- National Employment Survey (ENE) with paper questionnaire until 2019.
- Research: economic sector and/or occupation of the target population.
- Gloss: Description of the respondent's economic activity and/or occupation.
- Classification and coding of texts

The case of ENE

70,000 cases to be
classified per month

+3,000 hours of work
per month

Multiclassification
problems

- The coding team manually codes each survey according to the international classification. Classification categories determined by **International Classifications (UNSTAT, ILO)**



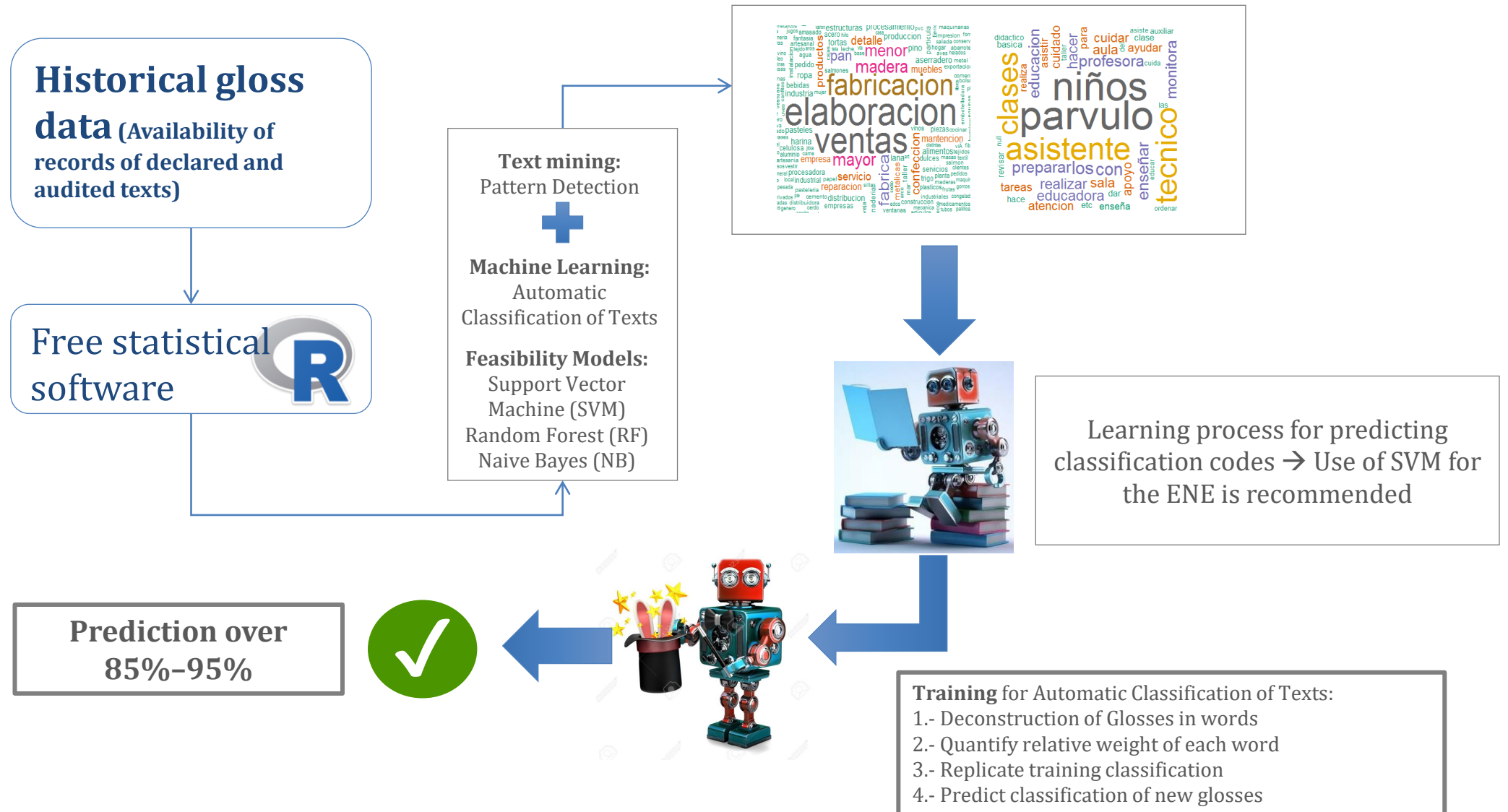
Survey response information
and coding are digitized.
(This task was done
manually)

Dimension of cases to be classified

-> **ENE (per month):**

- Target population: 23 thousand people approx.
- 5 open-ended questions on activity and occupation.
- Use of the international classifications.
 - International Standard Classification of Occupations (ISCO)
 - International Standard Industrial Classification of All Economic Activities (ISIC)
- **+ 70 thousand cases to be continually classified.**

Solution implemented in 2019



Results of the production process in 2019

COLLECTION

CODING

DATA ENTRY

CONSTRUCTION OF THE
DATABASE FOR TRAINING THE
MODEL

- Substitution of manual for automatic coding based on SVM methodology.
- Reduction of monetary costs and **working hours** for the ENE
- Implementation May 2019:
 - ISCO 08.cl series from the January–March 2017 quarter
 - CAENES (ISIC REV. 4.cl) series from the January–March 2013 quarter



Impact

- 1.- Shorter processing time.
- 2.- Lower labour cost.
- 3.- Greater precision.
- 4.- Higher quality of statistical information for decision making.

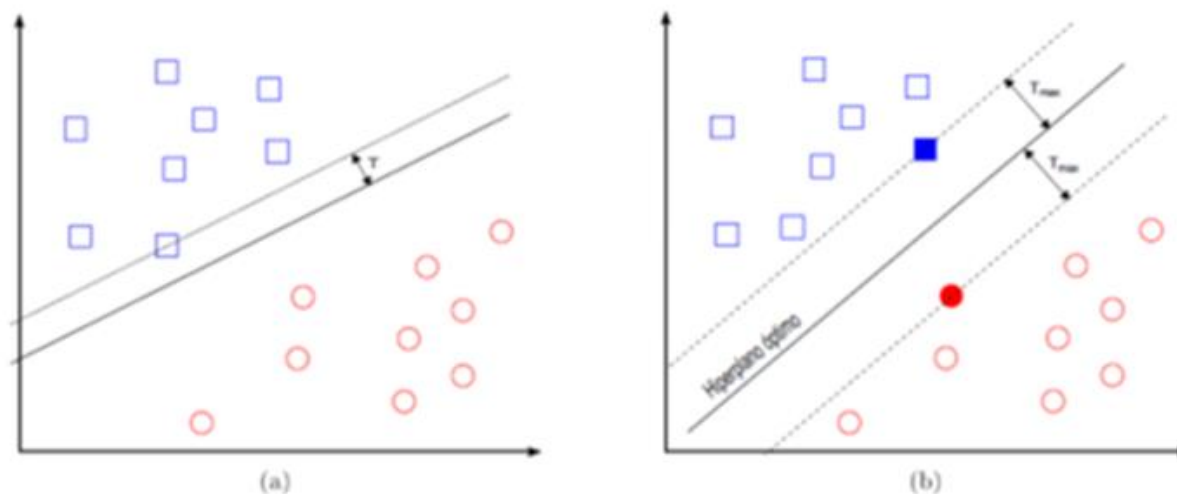
2. Methodological background

Method: SVM as an optimal method for gloss classification ENE, INE 2019.

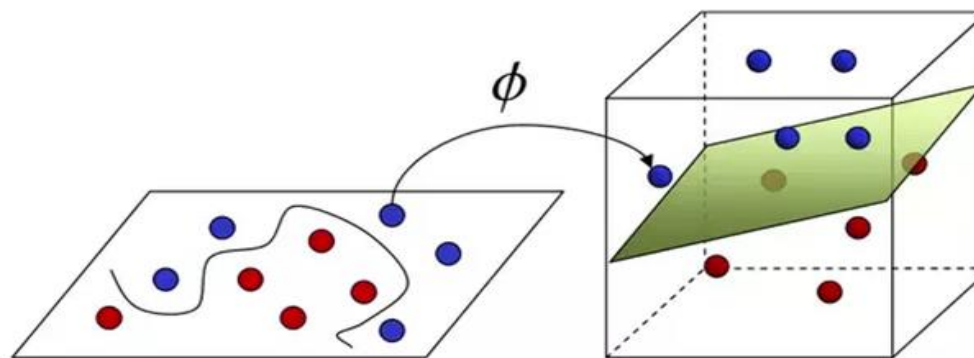
**SVM: Support
Vector machine**

**ENE → Glosses
2017 → +180
thousand records**

Linear separation of elements to be classified



Multidimensional separation of elements to be classified



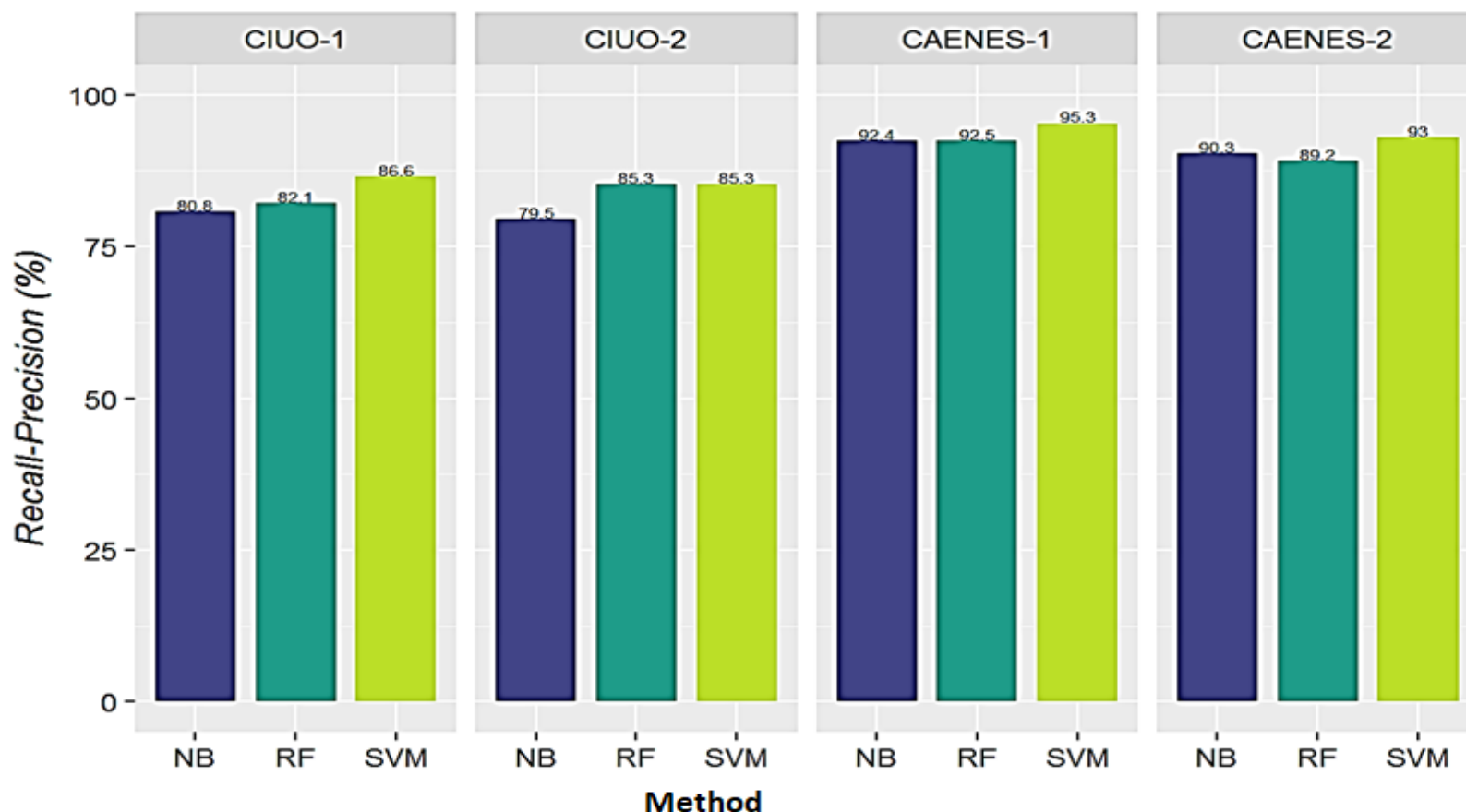
Matrix relative weights:

$$a_{ik} = f_{ik}$$

Based on the use of the matrix of relative weights of the words within the texts and within each category (sector or group), a mathematical rule (kernel) is applied to optimize the distance that separates the classification of a text between two categories.

Study determines SVM to be optimal method for ENE text classification

Recall-Precision* performance by method for ISCO-88 and CAENES Classifications



***Measurement of precision:**
Matched cases/Total cases

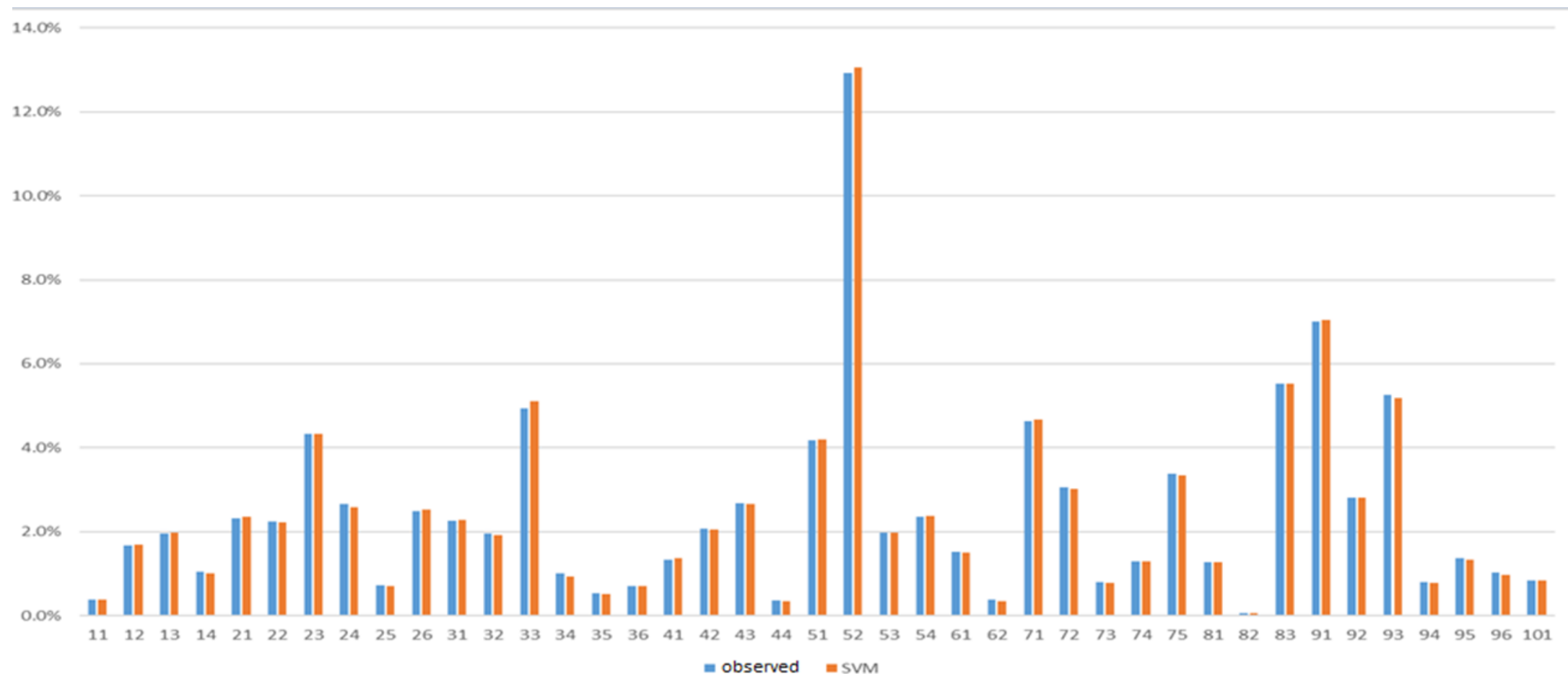
The model is trained with 80% of the data and the remaining 20% is evaluated. The *recall precision* measures the matches of the code assigned by SVM with the total number of codes in the training base.

SVM outperforms the others by up to 6 pp.

ENE adopts new technique

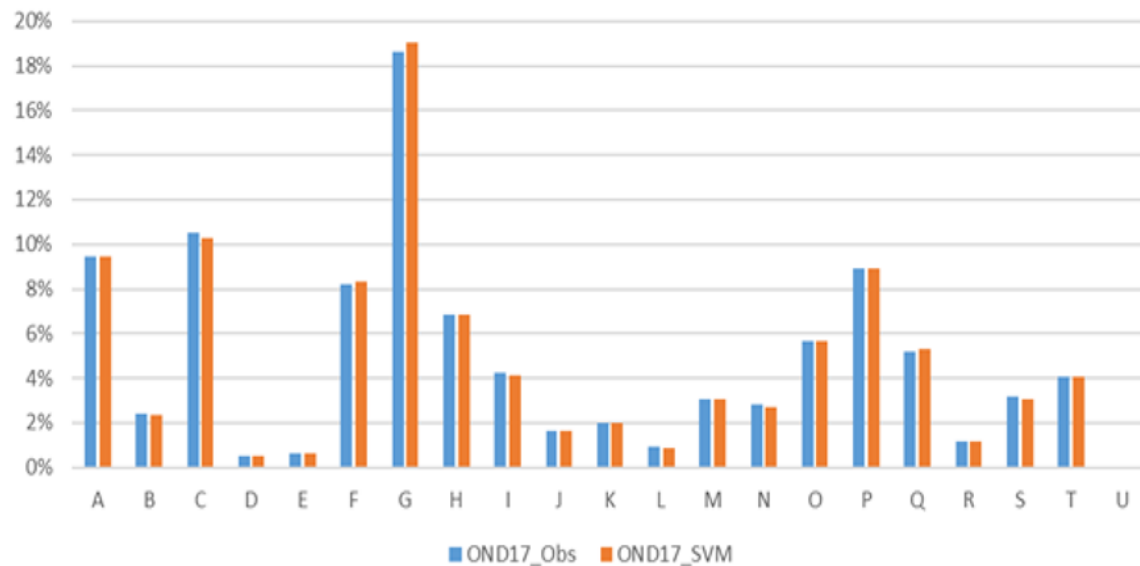
- Update training data to 2018
- Apply SVM process
- Determine ex-post analysis rules
- Develop training base for transition to ISCO 08.cl

Distribution % of glosses ENE 2017 coded manually (Observed) and with SVM.
ISCO 08.CL, with audit.

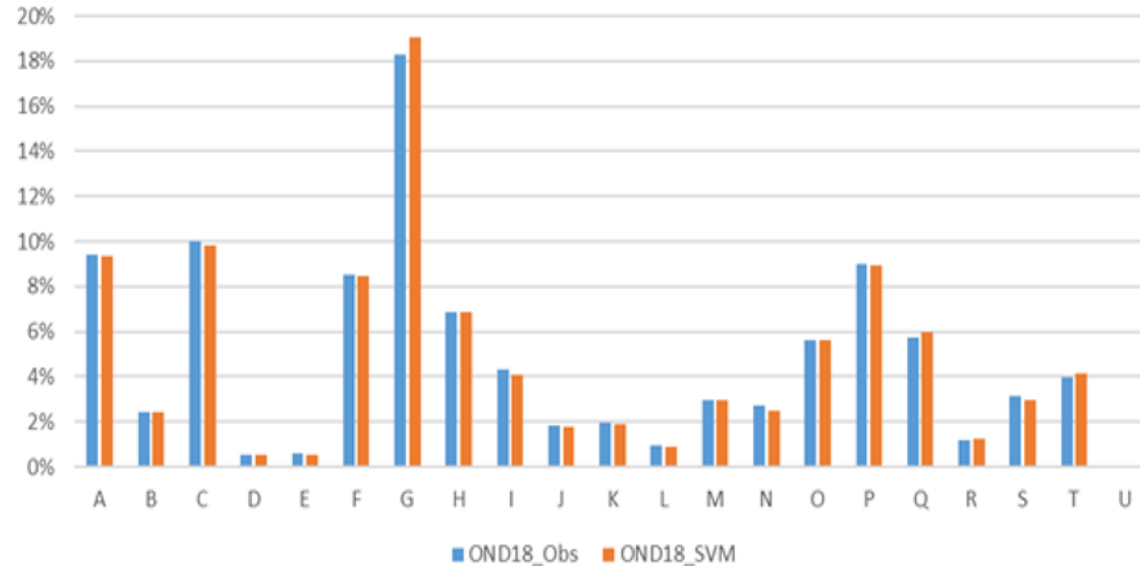


ENE Application: % distribution of estimated employed by branch ENE OND17–OND18

% Distribution of Estimated Employed (fact) according to economic sector
Manual (Obs.) and automatic (SVM) codification - ENE, OND 2017



% Distribution of Estimated Employed (fact) according to economic sector
Manual (Obs.) and automatic (SVM) codification - ENE, OND 2018



Main results:

The structure of distribution among the classification categories does not undergo major changes when switching to an automatic coding model.

There are changes in the economic variations between the two codifications. However, an analysis of confidence intervals shows that they overlap at the national and regional levels. The use of models helps in minimizing unsystematic errors derived from manual coding.

3. Automatic Classification Procedure for ENE Texts 2020

Implementation of the use of a mobile capture device.

National Employment Survey (ENE) open text classification

Field research information related to

Occupational group → ISCO 08.cl classification

Economic activity → CAENES classification (ISIC Rev. 4)

SECTION B

CHARACTERIZATION OF THE PRINCIPAL ACTIVITY

B1 What is the job, work, or occupation that you performed last week?


What tasks did you do in this occupation?

CIUO

Internal use

--	--	--	--


B13 What does the company, business, or institution that pays your wages do?

 The description refers to the Activity +
Good or Service + Prime Material or Type
of Sale


CAENES

Internal use


--	--	--	--

 Independent Workers (If B2=1 or B3=2, 3, or 4)

B14a What do you do as an own-account worker?

 Wage earners (If B2=2, B3=1, or B6=3)

B14b What does the company, business, or institution where you work do?

 The description refers to the Activity +
Good or Service + Prime Material or Type
of Sale

CAENES

Internal use

--	--	--	--

Automatic Text Classification Procedure ENE 2020 to date

Digital collection (DMC)

- Continuous survey conducted with mobile computing device (Tablet)
- Daily synchronization of collected cases
- Data storage in SQL server



Data extraction from SQL server (JSON)

- The data collected to date is extracted weekly
- SQL server access → Extract JSON → Transformation to table for ID information and required texts



Execution SVM Models

- Cleandata procedure for available texts
- Upload model and corresponding term document matrix (CAENES / CIU008.cl)
- Execute predict function
- Stores encoded cases on server



Quality control

- Post SVM critical case detection algorithm
- Report generation for manual review
- Manual review of identified cases
- **Storage of reviewed cases → Table with person identifier (ID) + audited code**



ENE database integration

- Procedure consolidates codes determined by the SVM model with the manual audit completed
- Storage for ENE database production (Short-term)
- Storage for training base update (Structural)

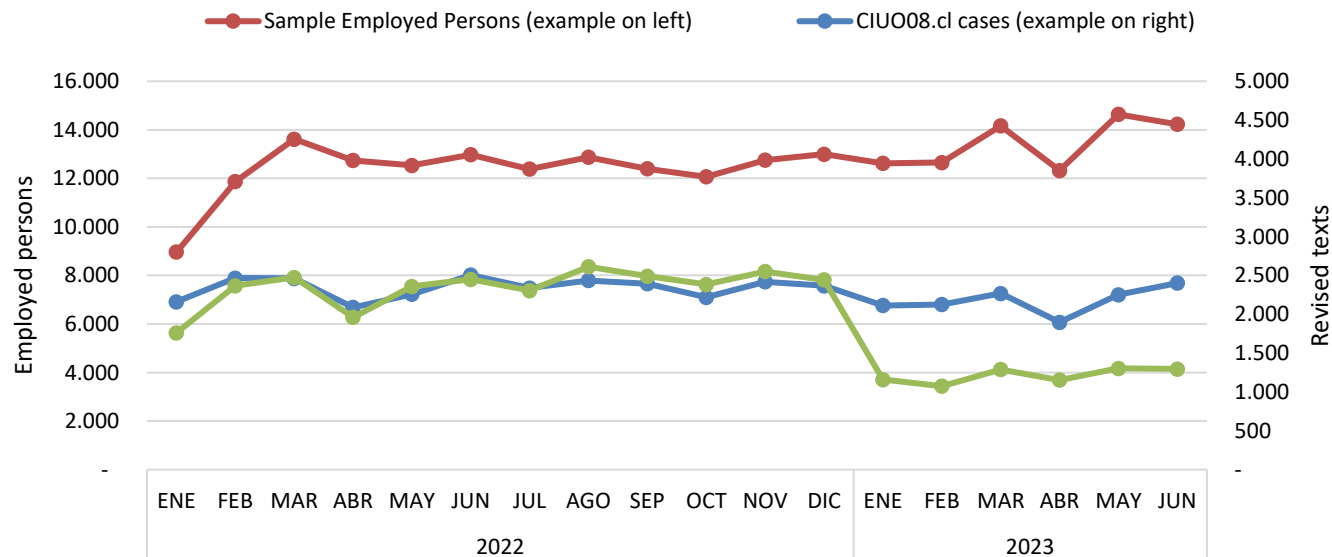


4. Quality control

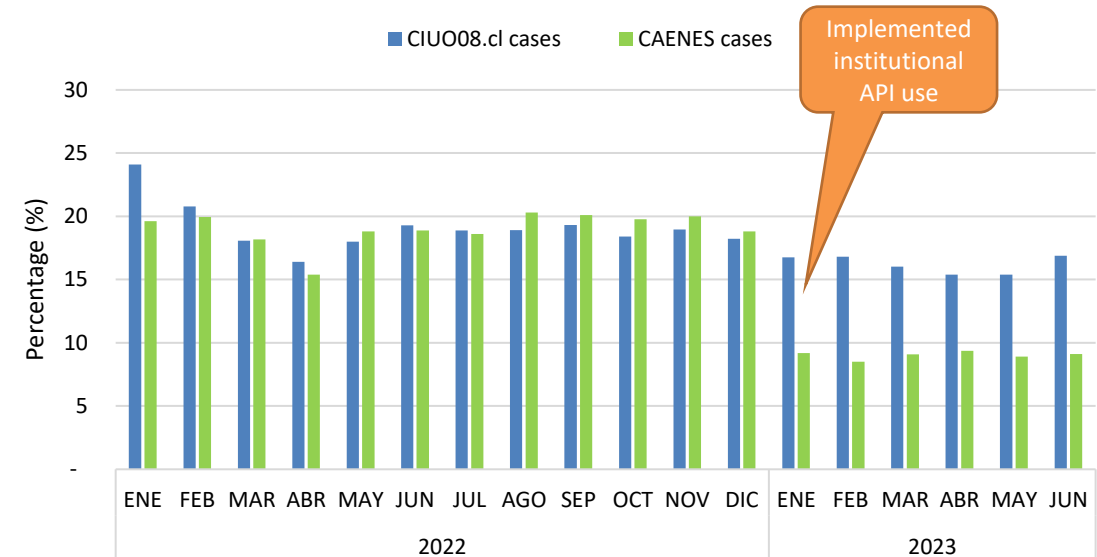
Post SVM algorithm detects

- Classification with 2-digit model without correspondence with 1-digit model
- Differing classification between SVM and institutional API implemented in 2023
- Critical cases identified with rules determined from the prevalence of keywords (e.g., CAENES, Construction + Welding). Example ISCO, secretary vs. executive secretary
- Manual review by an expert in classifications CIUO08.cl and CAENES
- Institutional API is used to resolve inconsistencies between 2-digit and 1-digit classification, reducing the number of cases to be reviewed by the expert.

Evolution of the sample of employed persons and cases reviewed manually by CIUO08.cl and CAENES. January 2022 to June 2023



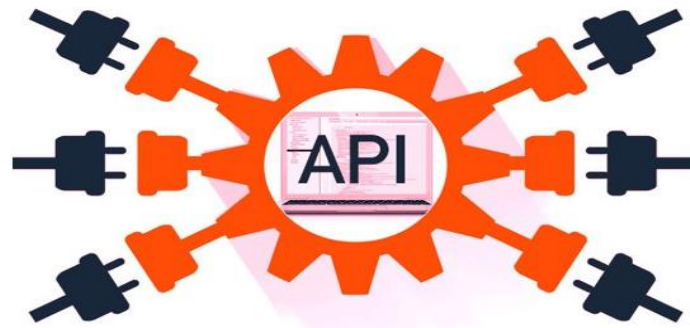
Evolution of the percentage of cases reviewed compared to the total employed population. Jan 2022 to Jun 2023



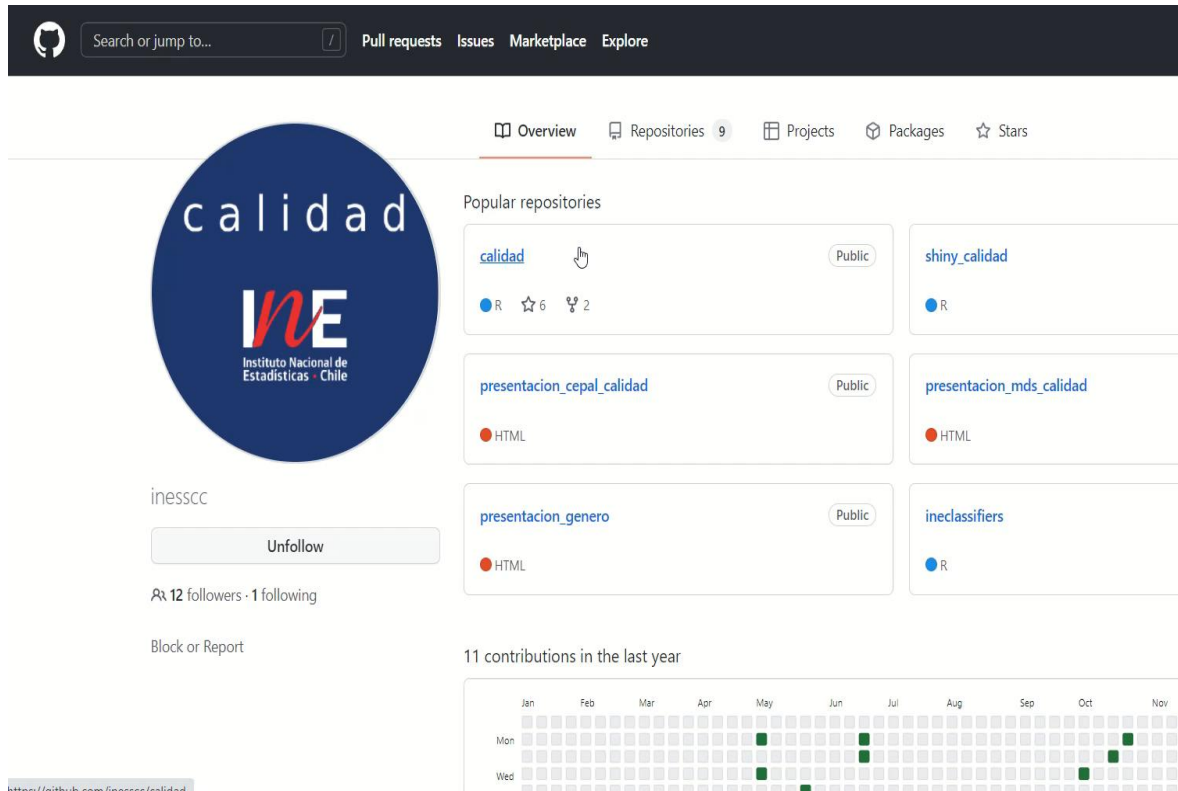
5. Automatic coding API service for occupation and economic activity

- It arises from the need for various statistical operations to have a “**generalist**” **algorithm** for the coding of occupation and branch of economic activity, that is, one that depends on the smallest number of variables possible.
- Training bases were generated with special emphasis on the quality of the classification rather than the quantity of labeled glosses (90% cross-coding and expert review of discrepancies).
- A **neural network** model was trained that uses **Gated Recurrent Unit and Word Embeddings**
- Tutorial on the institutional website: <https://www.ine.gob.cl/calidad-estadistica/clasificaciones/api-codificacion>

- A difficulty was detected in making the trained model available for use in statistical operations.
- The problem was the difficulty of getting R and Python **dependencies setup** on local computers.
- The solution was to provide the service as a API Rest from an institute server which can be consulted from various programming languages.



Automatic coding API service



The screenshot shows the GitHub repository page for 'calidad' by 'inssc'. The repository is public and has 6 stars and 2 forks. It is written in R. The page also shows a list of popular repositories, including 'shiny_calidad', 'presentacion_cepil_calidad', 'presentacion_mds_calidad', 'presentacion_genero', and 'ineclassifiers'. A contribution graph shows 11 contributions in the last year.

Resultados CAENES 2 dígitos

modelo	acc	macro	micro	weighted
seq_2d	0.9075	0.7536	0.9075	0.9083
tfidf_2d	0.9076	0.7579	0.9076	0.9081
emb_simple_2d	0.9021	0.7631	0.9021	0.9039
emb_gru_2d	0.9048	0.7630	0.9048	0.9052

Resultados CIUO 2 dígitos

modelo	acc	macro	micro	weighted
seq_2d	0.8456	0.7249	0.8456	0.8476
tfidf_2d	0.8412	0.7355	0.8412	0.8431
emb_simple_2d	0.8324	0.7220	0.8324	0.8348
emb_gru_2d	0.8526	0.7364	0.8526	0.8543

A short video tour of the github repository of the coding service is shown (it is public), and the performance of 2-digit CAENES and ISCO in the chosen models is also highlighted.

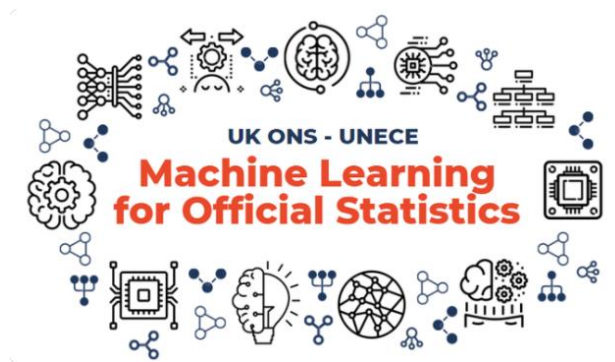
Repository in github:

<https://github.com/inssc/ineclassifiers>



Automatic coding API service

Closing conference of the year ML Group – UNECE 2021



Automatic coding API demo

```
RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function | Addins
Untitled1* x
Source on Save
1 |
2 # Predecir a un dígito
3 library(httr)
4 library(feather)
5 caenes <- read_feather("C:/Users/Ignacio/Downloads/test.feather")
6 request <- httr::POST("http://143.198.79.143:8080/predict",
7                       encode = "json",
8                       body = list(text = caenes$glosa_caenes[1:5],
9                                   classification = "caenes",
10                                   digits = 1)
11 )
12
13 caenes$glosa_caenes[1:5]
14
15 httr::status_code(request)
16 response <- httr::content(request)
17 response
```

```
1:1 (Top Level)
Console Terminal Jobs
R 4.1.0 ~/
> |
```

INE Chile presented its experience at the closing of the annual Machine Learning Conference in 2021.

A brief demonstration of an API query from R is shown.

6. Next steps

1. Updating training bases
2. Feasibility study on the transition to Deep Learning methodology
 - a. Collaboration Agreement, Master in Data Science, Universidad de Chile
 - b. Exploration and use of Transforms
3. Updating SVM Models
4. Evaluation model to be used
5. Evaluation of possible rectification of published series



Thank you

NATIONAL STATISTICS INSTITUTE