Work Package 10
6th November, 2024

Introducing CodIA

Codifying economic activities with Fasttext

Adrián Pérez Bote

Head of classifications unit, Statistics Spain

INE
Instituto Nacional de Estadística

Introduction: CNAE, Spanish version of NACE

CNAE

Clasificación Nacional de
Actividades Económicas

INE

Instituto Nacional de Estadística

Introduction:
CNAE, Spanish version of NACE

What is CNAE?

Clasificación Nacional de Actividades Económicas, resulting from adapting the corresponding European version, the NACE.

The objective of this classification is to establish a hierarchical set of economic activities that can be used to:

- Favor the implementation of national statistics that can be differentiated according to the established activities.
- Classify statistical units and entities according to the exercised economic activity.

Instituto Nacional de Estadística

Introduction:
CNAE, Spanish version of NACE

CNAE 2025

- Version of the CNAE effective January 1, 2025
- Replaces CNAE 2009
- Adapts NACE Rev 2.1. to the national context
    - Up to the third level of disaggregation, they are identical.
    - 10 more classes are added by splitting 10 NACE Rev 2.1 classes.

Defining the coder: CodIA

Defining the coder: CodIA

- We propose the development of an automatic coder using ML techniques and, specifically, NLP.
- This coder would mark an evolution of the current tools based on heuristics (AYUDACOD and AUTOCOD).
- Objective: from a description of the economic activity, return the corresponding CNAE25 class.

Defining the coder: CodIA

- Two use cases:
  - Automatic coding
  - Coding assistance
- Two functionalities:
  - Coding from a literal
  - Coding from a literal and the CNAE09 class.
- Two access modes:
  - Web service
  - Interface
- The output is a list of possible classes, each of which contains:
  - A code
  - A title
  - Explanatory notes (includes and excludes)
  - A score

Defining the coder: CodIA

Toolbox

- Python: programming language
- Anaconda: AI platform
- Jupyter notebooks: interactive computing
- Fasttext: Python library for NLP classifier development
- Python libraries: Numpy, Pandas, Matplotlib, Plotly…
- Excel: exploratory analysis and dataset labeling

Introducing Fasttext

Instituto Nacional de Estadística

Introducing Fasttext

- **Open source** NLP library developed by Facebook AI Research in 2016.
- Available for **Python** and command line
- Designed for:
  - Supervised learning: **text classification**.
  - Unsupervised learning: n-dimensional vector space word representations
- **Very fast** to train and test, in contrast to deep learning networks, Transformers and LLMs (state of the art)

Introducing Fasttext

Some keys

- It represents each word by a low-dimensional vector.
- Fasttext actually includes representations for n-grams of words, instead of “single” words, in order to take into account the order in which they appear locally.
- It also associates a vector to a text, obtained by summing the vectors of the n-grams that make it up.

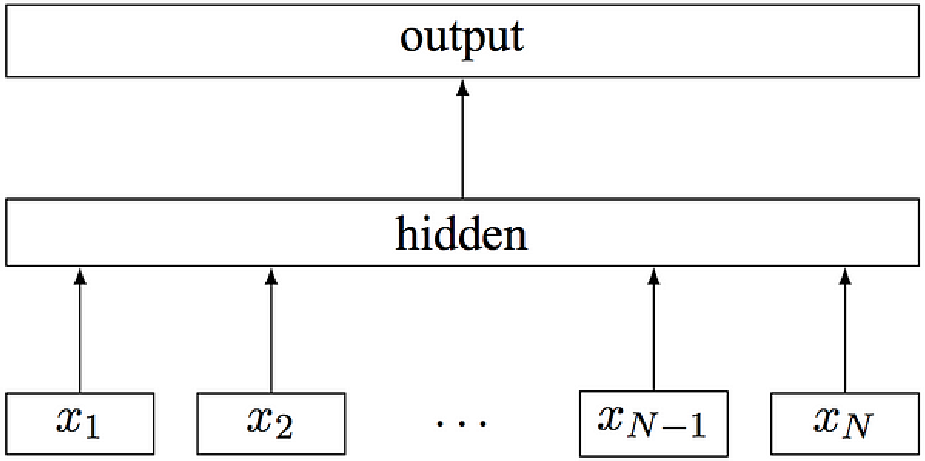


Figure 1: Model architecture of `fastText` for a sentence with N ngram features $x_1, \ldots, x_N$. The features are embedded and averaged to form the hidden variable.

Datasets: real data

Instituto Nacional de Estadística

Datasets: real data

Surveys (WIP, 100K samples)

- Structural Business Survey
- Research and Development Survey
- Ad-hoc survey taking place in Q3 & Q4

SBS + Business register(600K samples)

- We cross the two data sets
- Social purpose, not exactly economic activity descriptions

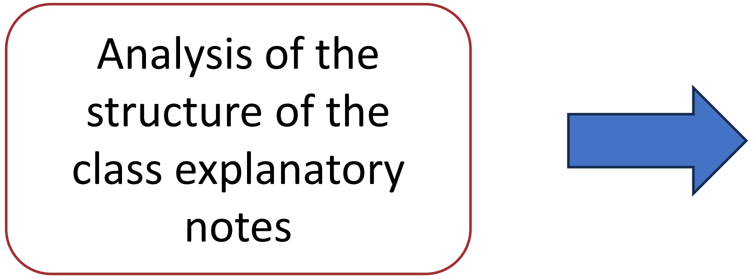
Datasets: synthetic data

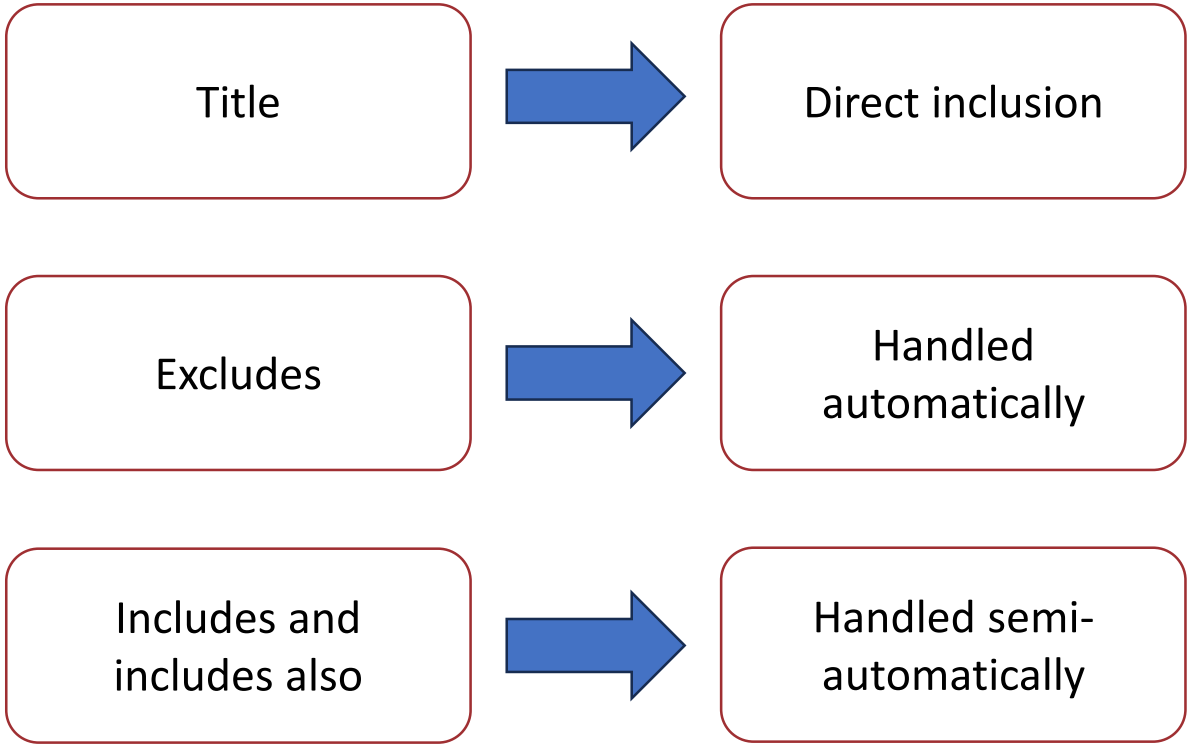
Datasets: synthetic data

Why add synthetic data?

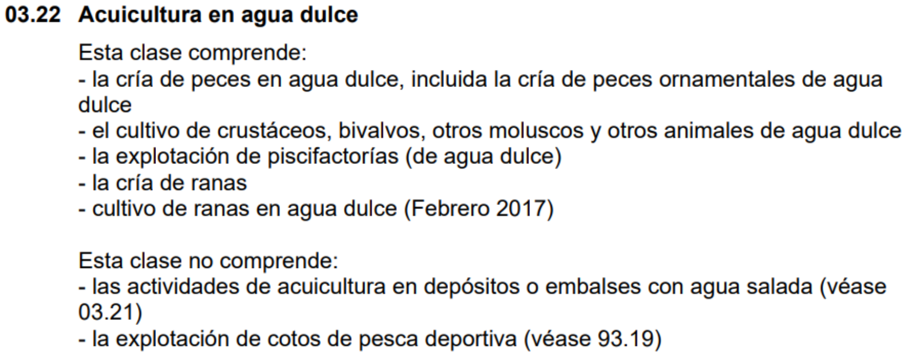
- Ensure a reliable training dataset
- Ensure a base syntax and glossary for each class
- Include data for those classes not represented in the rest of the dataset
- Train models for categories for which no class data is available

Datasets: synthetic data





Fuente: Notas CNAE2009



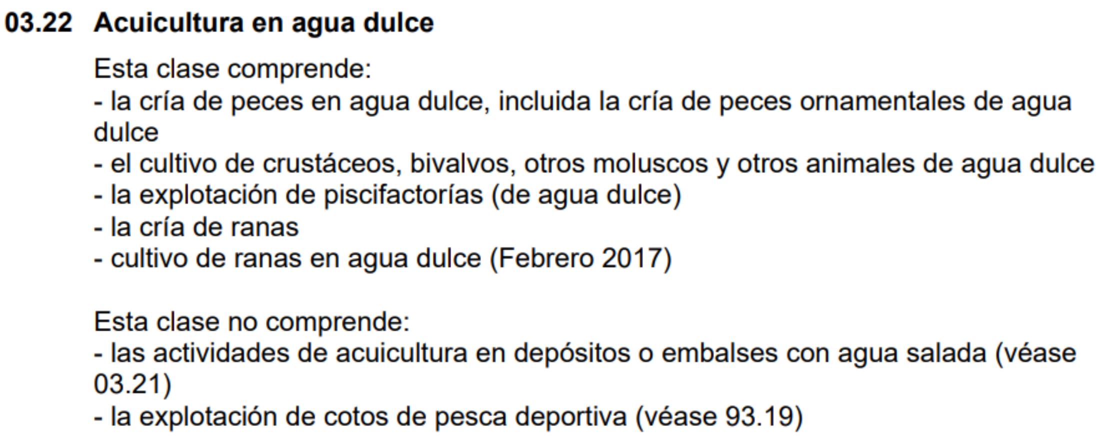
03.22 Acuicultura en agua dulce

Esta clase comprende:
- la cría de peces en agua dulce, incluida la cría de peces ornamentales de agua dulce
- el cultivo de crustáceos, bivalvos, otros moluscos y otros animales de agua dulce
- la explotación de piscifactorías (de agua dulce)
- la cría de ranas
- cultivo de ranas en agua dulce (Febrero 2017)

Esta clase no comprende:
- las actividades de acuicultura en depósitos o embalses con agua salada (véase 03.21)
- la explotación de cotos de pesca deportiva (véase 93.19)

Instituto Nacional de Estadística

Datasets: synthetic data



03.22 Acuicultura en agua dulce

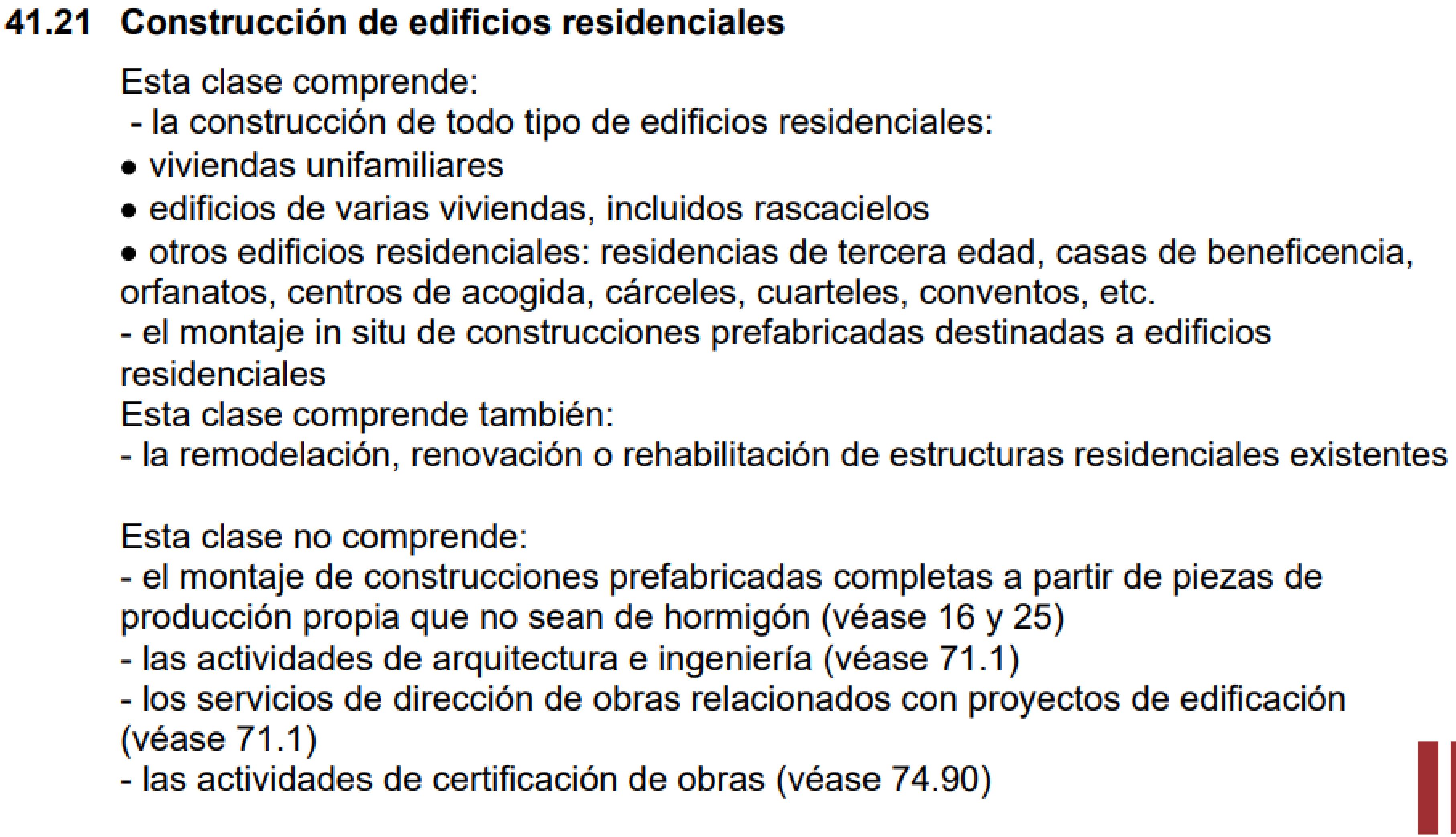
Esta clase comprende:
- la cría de peces en agua dulce, incluida la cría de peces ornamentales de agua dulce
- el cultivo de crustáceos, bivalvos, otros moluscos y otros animales de agua dulce
- la explotación de piscifactorías (de agua dulce)
- la cría de ranas
- cultivo de ranas en agua dulce (Febrero 2017)

Esta clase no comprende:
- las actividades de acuicultura en depósitos o embalses con agua salada (véase 03.21)
- la explotación de cotos de pesca deportiva (véase 93.19)

COD_CNAE	Literal
0322	Acuicultura en agua dulce
0321	Las actividades de acuicultura en depósitos o embalses con agua salada
9319	La explotación de cotos de pesca deportiva
0322	La cría de peces en agua dulce
0322	La cría de peces ornamentales en agua dulce
0322	El cultivo de crustáceos
0322	El cultivo de bivalvos
0322	El cultivo de otros moluscos
0322	El cultivo de otros animales de agua dulce
0322	La explotación de piscifactorías de agua dulce
0322	La cría de ranas
0322	Cultivo de ranas en agua dulce

Title
Excludes
Compounded includes (with “y” or “,”)
Simple included

Datasets: synthetic data



41.21 Construcción de edificios residenciales

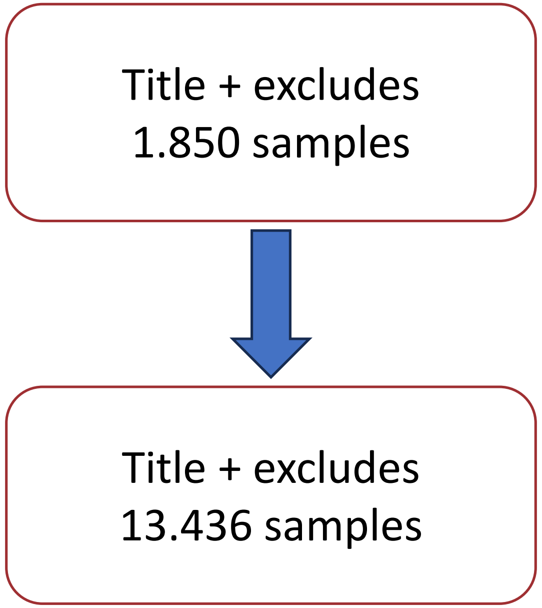
Esta clase comprende:
- la construcción de todo tipo de edificios residenciales:
● viviendas unifamiliares
● edificios de varias viviendas, incluidos rascacielos
● otros edificios residenciales: residencias de tercera edad, casas de beneficencia, orfanatos, centros de acogida, cárceles, cuarteles, conventos, etc.
- el montaje in situ de construcciones prefabricadas destinadas a edificios residenciales
Esta clase comprende también:
- la remodelación, renovación o rehabilitación de estructuras residenciales existentes

Esta clase no comprende:
- el montaje de construcciones prefabricadas completas a partir de piezas de producción propia que no sean de hormigón (véase 16 y 25)
- las actividades de arquitectura e ingeniería (véase 71.1)
- los servicios de dirección de obras relacionados con proyectos de edificación (véase 71.1)
- las actividades de certificación de obras (véase 74.90)

Datasets: synthetic data

- Use of a basic thesaurus (dictionary of synonyms) to enrich the data set

```
{
        'la fabricación':sin_fabricacion,
        'las actividades': sin_actividades,
        'el comercio': sin_comercio,
        'la producción': sin_produccion,
        'la reparación': sin_reparacion,
        'los servicios': sin_servicios,
        'los productos': sin_productos,
        'el cultivo': sin_cultivo
}
```



Title + excludes
1.850 samples

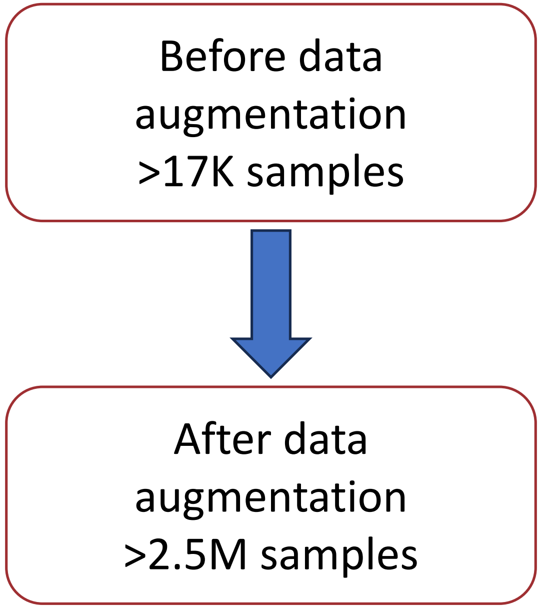
Title + excludes
13.436 samples

TOTAL: 19.233 samples

Datasets: synthetic data

Data augmentation

- Use of vocabulary tables and synonyms from an INE tool to assist in classification



Before data augmentation >17K samples

After data augmentation >2.5M samples

TOTAL: >2.5M **samples**

Ensuring data quality

Ensuring data quality

Training and test datasets

- We preprocess the text by converting to lowercase, removing special characters, numbers, etc.
- Synthetic data adds volume to all classes, which is especially relevant for minority classes.
- At the same time, we ensure perfect labeling.
- The ad-hoc survey ensures high quality data, as companies only need to focus on answering a few questions about their CNAE.
- We eliminate erroneous training patterns in the preprocessing phase.
- We measure the quality of the actual data by applying manual labeling to a sample of each set.
- The test set is reviewed manually, in detail.
- Each subset (real or synthetic) is evaluated by adding and removing it from the training set and obtaining performance metrics.

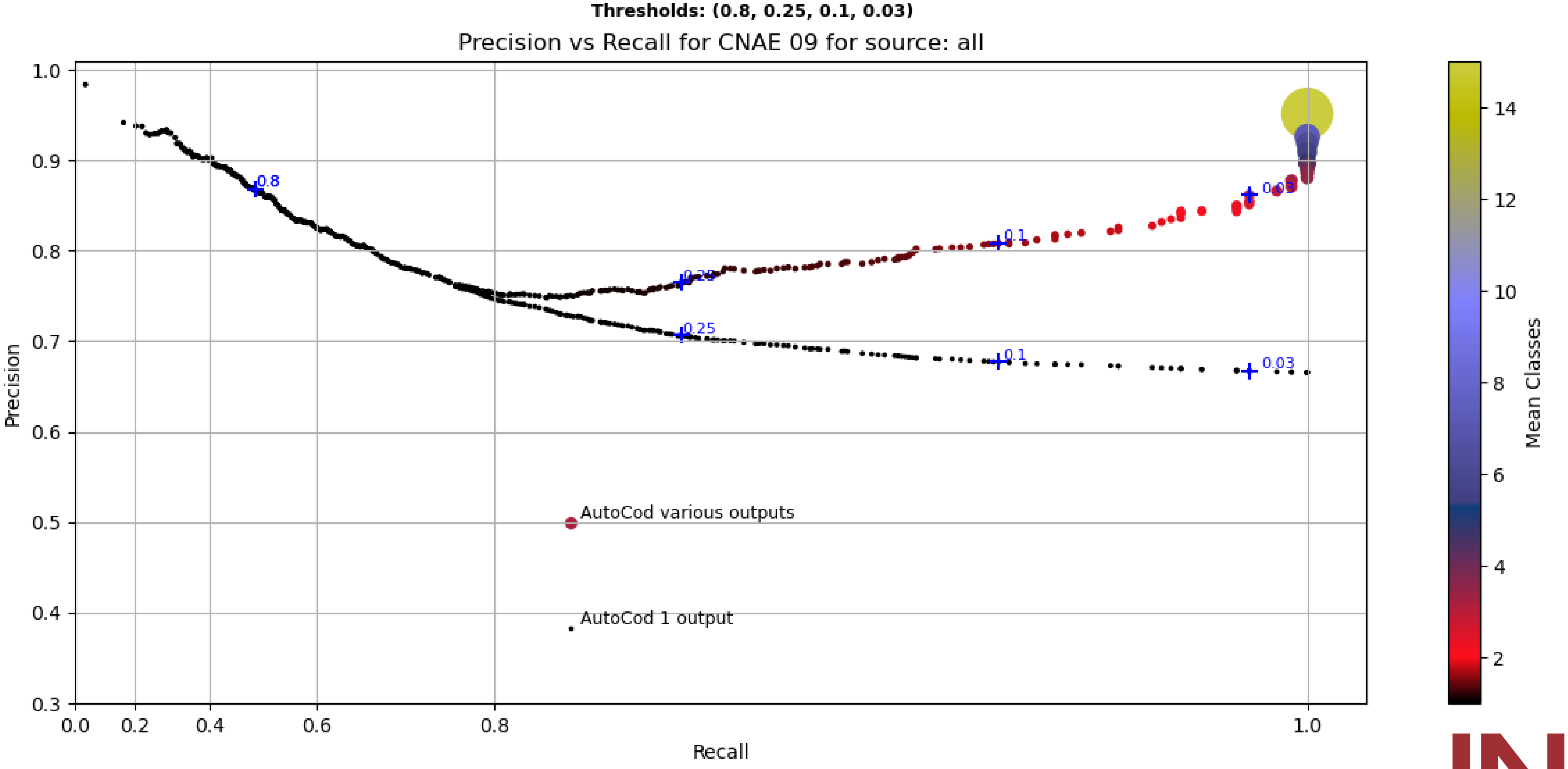
Ensuring data quality

Input data in production

- We have developed an additional layer to CodIA
- It consists on a binary Fasttext classifier that discriminates between economic activity descriptions and any other texts
- The CNAE codifier is executed only if the predicted class is positive
- It performs up to 99% accuracy

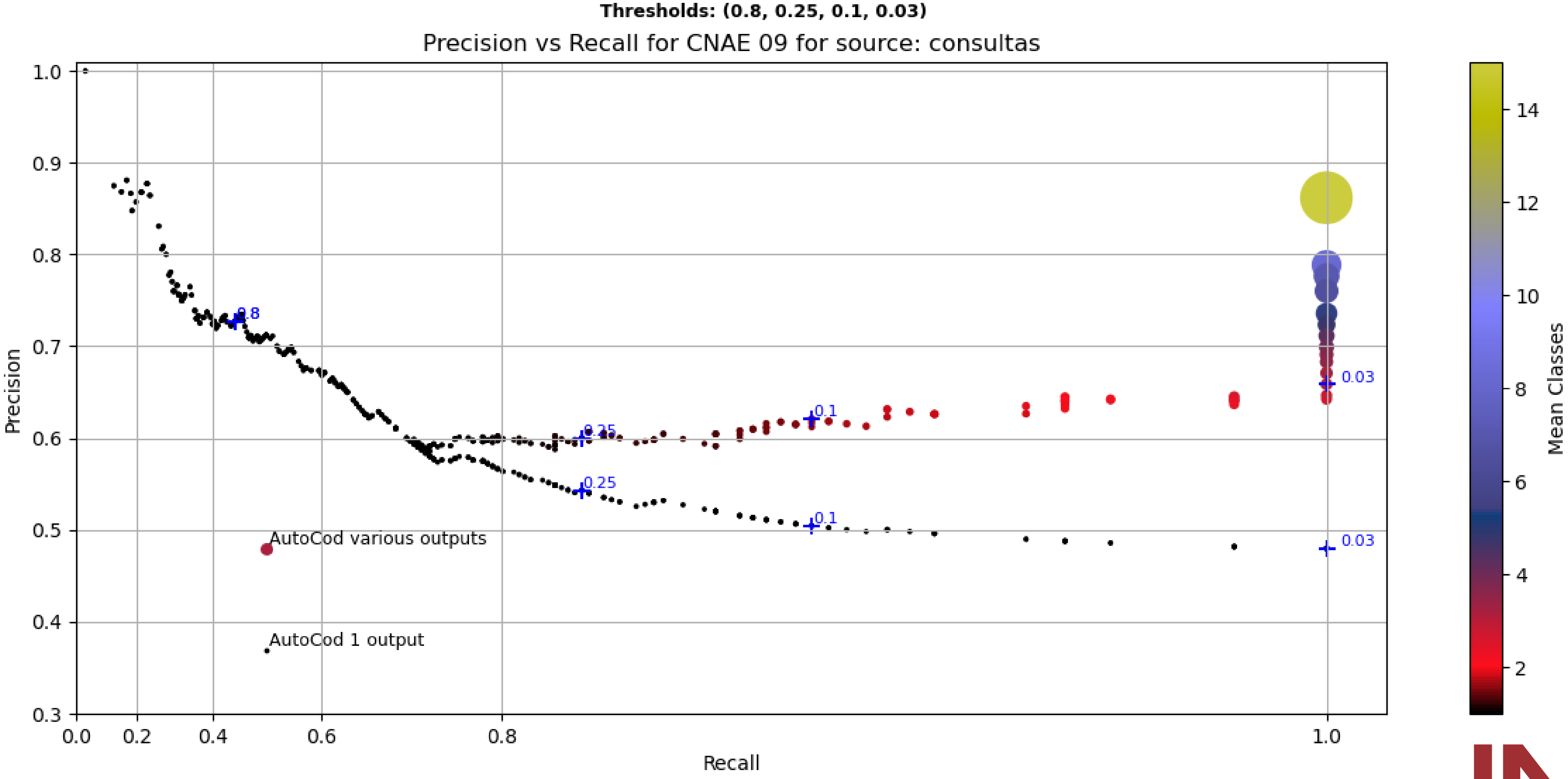
Results

Performance



Instituto Nacional de Estadística

Performance



Instituto Nacional de Estadística

Live demo!

Thank you for your attention
&
Let’s keep up codifying!

🥰

Contact

https://www.ine.es/

https://twitter.com/es_ine

https://www.instagram.com/es_ine_/

https://es.linkedin.com/company/ine-es

adrian.perez.bote@ine.es

carlos.saez.calvo@ine.es

Instituto Nacional de Estadística