Coding occupations in the Population and Housing Census of Ecuador using machine learning techniques

National Institute
of Statistics and Census
of Ecuador

Méndez Diana, Espinoza Victor

Table of contents

Introduction

International Standard Classification of Occupations ISCO – 08
(4th level)

→

National Standard Classification of Occupations
(6th level)

In your main job, what is the occupation or tasks you perform?

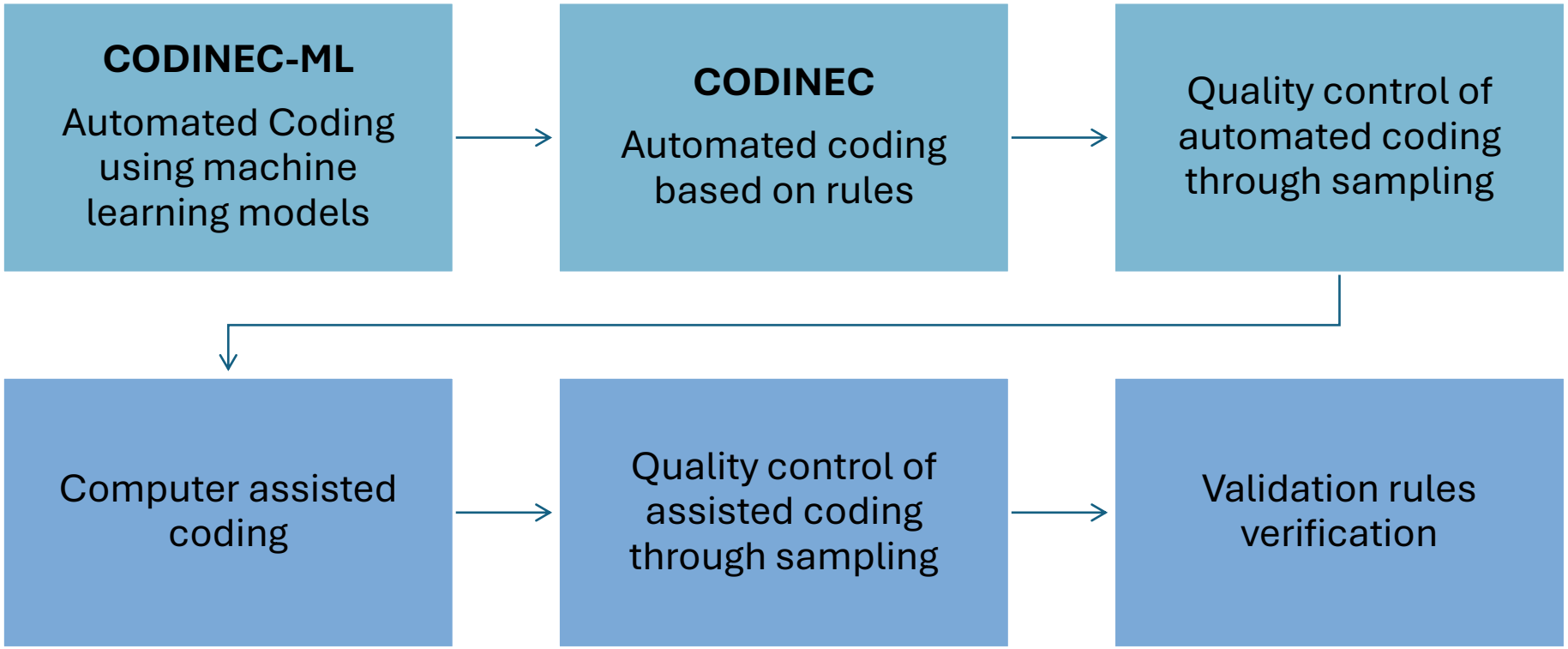
Coding occupations →

Occupation response

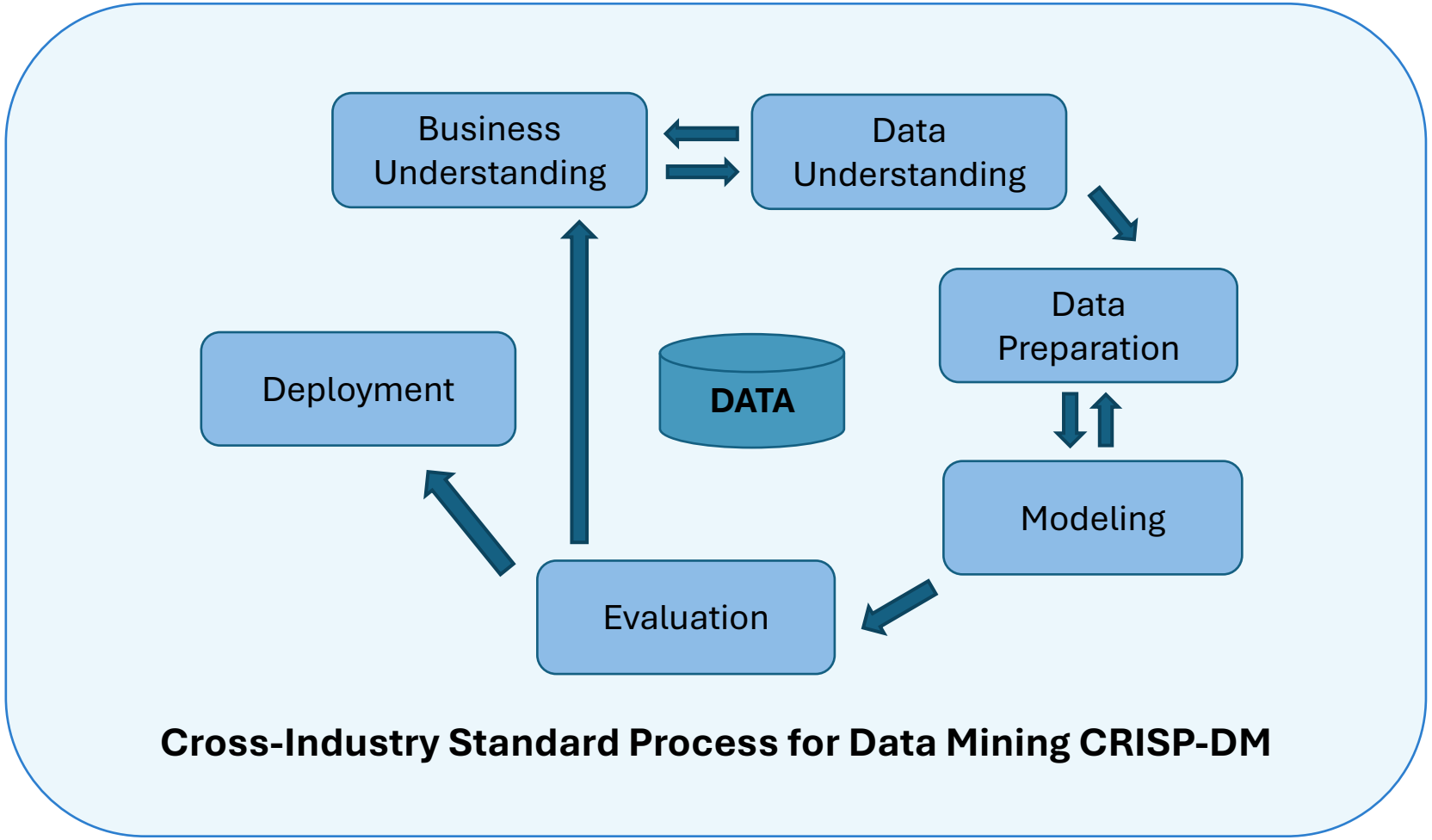
Ancillary information:

- Industry description
- Occupational category
- Age
- Educational level
- Social security

Coding process



Methodology for machine learning projects



Cross-Industry Standard Process for Data Mining CRISP-DM

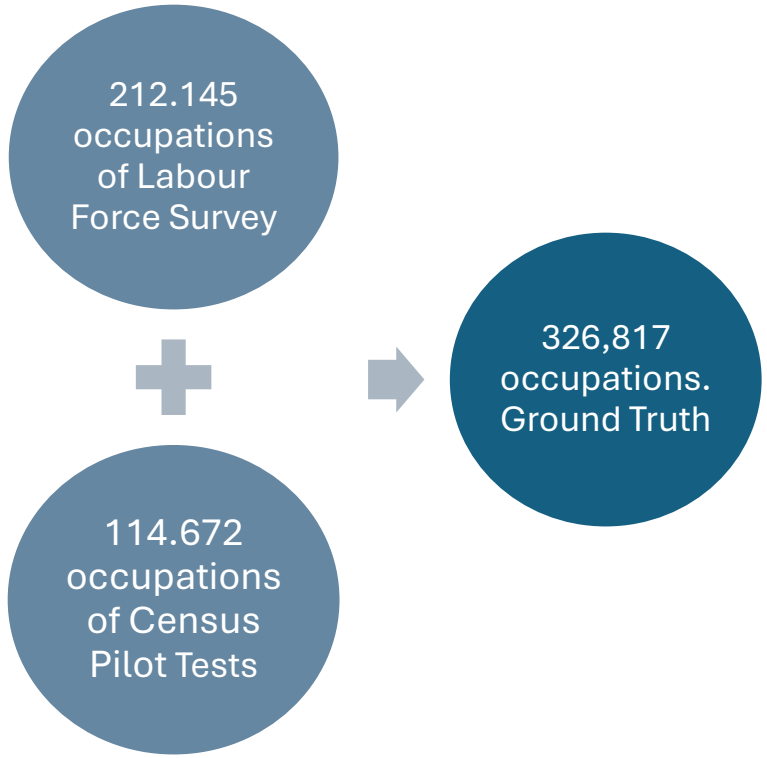
Developing machine learning models

Business Understanding

Coding ≈ 5.4 millon of occupations

Main concerns:	Recruiting , training and managing hundreds of coders.
	Ensuring sufficient financial resources for coding staff and operations.
	Maintaining consistent quality control.
Solution:	Automating a part of the process using machine learning techniques.

Data Understanding

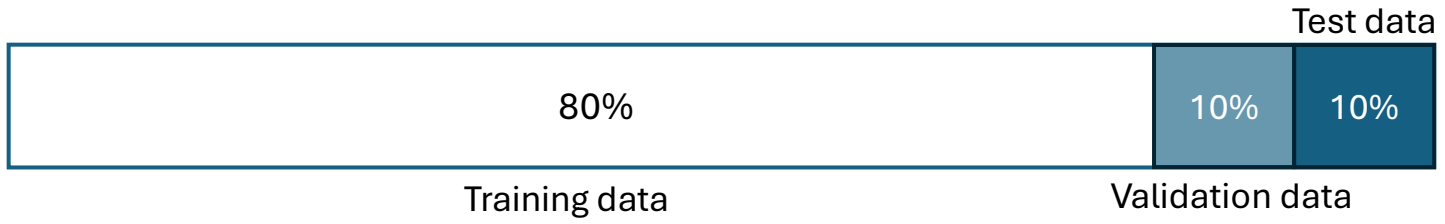


Developing machine learning models

Data Preparation

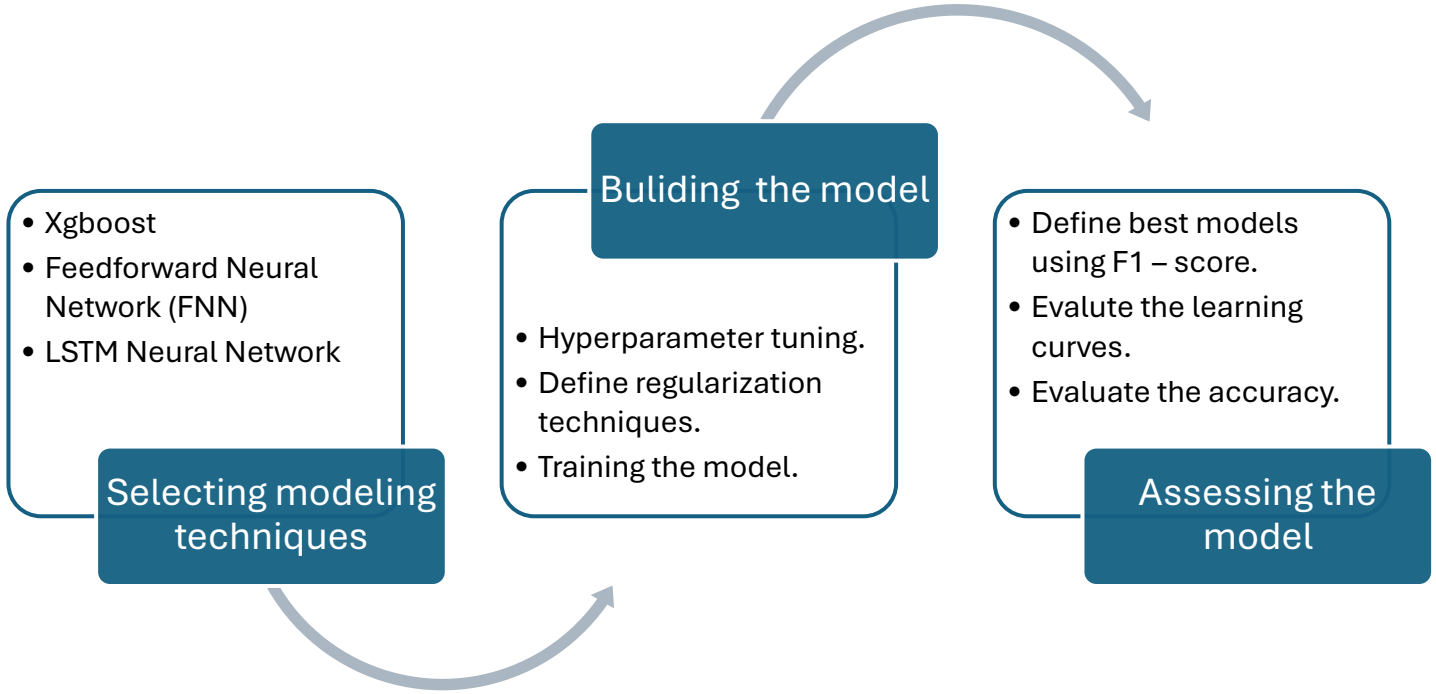
Text preprocessing techniques	Feature extraction techniques	Ancilliary data preprocessing	Data balancing
• Lowercasing • Text cleansing: Removing punctuation marks, numbers, extra spaces and accents. • Tokenization. • Removing stop words • Stemming.	• TF – IDF (Xgboost) • Word Embeddings (Artificial Neural Networks)	• One Hot Encoding	• Resample the training dataset (344 classes): under-sampling and over-sampling

Spliting data
(Holdout Validation)



Developing machine learning models

Modeling



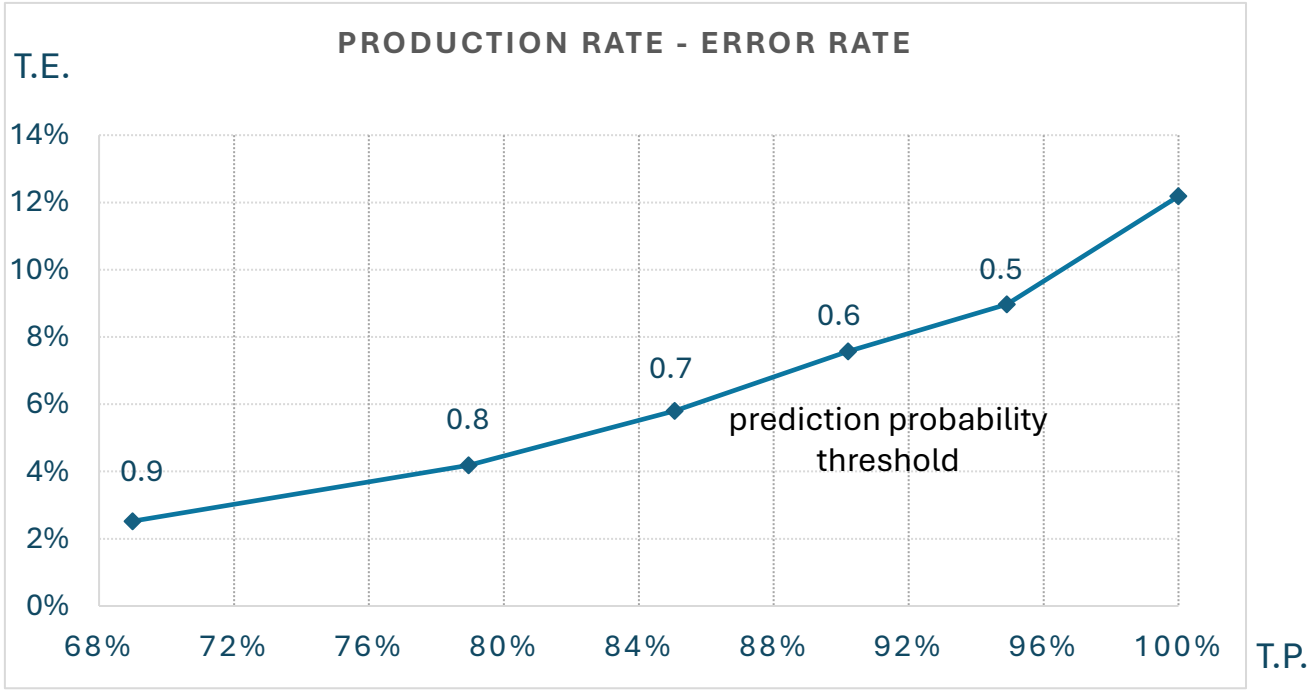
Developing machine learning models

Evaluation

Best model: **Feedforward Neural Network**

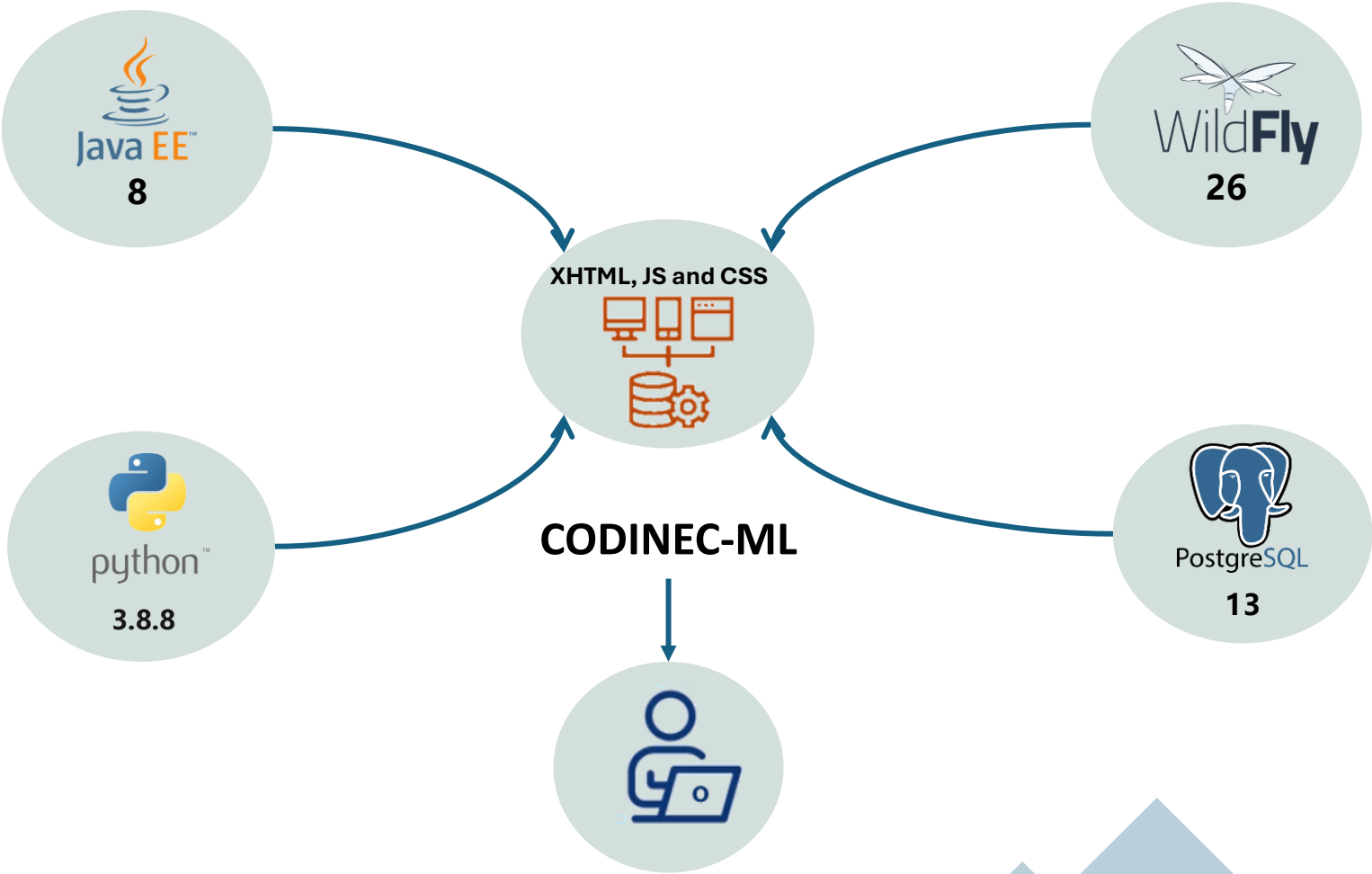
Accuracy: **87.82%**

Training time: ≈ **12 minutes**



Deployment in an integrated system

Web development tools and server



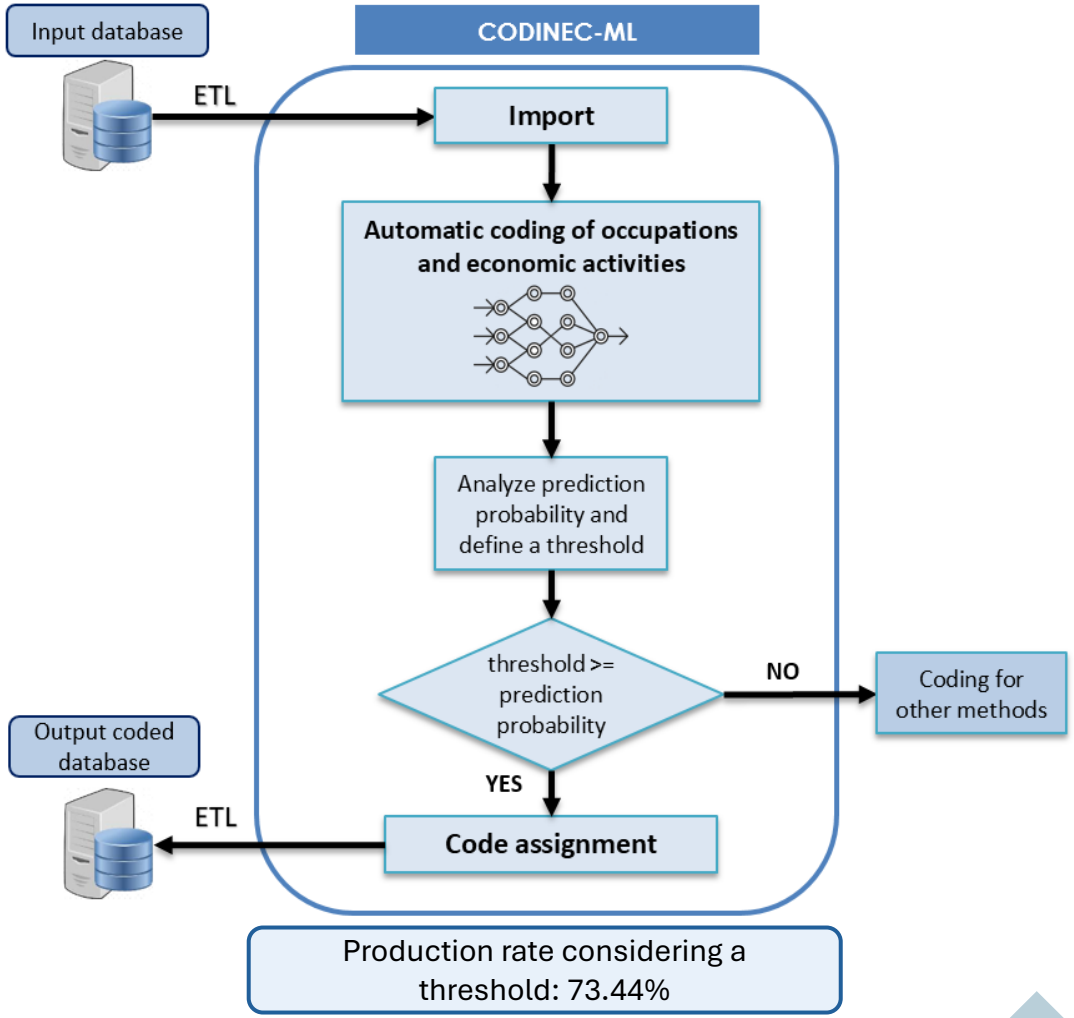
Deployment in an integrated system

Infrastructure Requirements

Feature	Database Server	Application Server	Web Server
Operating System	CentOS 7	CentOS 7	CentOS 7
CPU	24 cores	24 cores	24 cores
Memory	64 GB	32 GB	32 GB
Storage	1 TB	500 GB	500 GB

Deployment in an integrated system

Functionality



Key Takeaways

1. The base of any machine learning model lies in the training data. Ensuring the accuracy of labels is essential, as poor-quality ground truth can lead to unreliable predictions.

2. No model is perfect. However, before deploying a model in production, it must be considered that the error rate of the model is lower than of manual coding.

3. The production rate must be carefully calibrated, balancing available resources with the margin of expected error.

4. During the coding process, having a dedicated quality control team is non-negotiable. They play a critical role in monitoring the model's performance due to ambiguous and vague inputs that can affect the predictions.

Thank you for your attention!

INEC | Good numbers, better lives

@InecEcuador

@ecuadorencifras

@ecuadorencifras

INECEcuador