UNIVERSITY
INSTITUTE
OF LISBON

Artificial Intelligence applied to public employment data in the Portuguese Education sector

Inês Colaço Ascenso

Master in Computer Engineering

Supervisor:
PhD *Luís Miguel Martins Nunes*, Associated Professor,
Iscte - Instituto Universitário de Lisboa

Co-Supervisor:
PhD *Pedro Lopes da Silva Mariano*, Integrated Researcher,
ISTAR-Iscte - Information Sciences, Technologies and
Architecture Research Centre

September, 2025

[This page is intentionally left blank.]

TECHNOLOGY
AND ARCHITECTURE

Department of Information Science and Technology

Artificial Intelligence applied to public employment data in the Portuguese Education sector

Inês Colaço Ascenso

Master in Computer Engineering

Supervisor:
PhD *Luís Miguel Martins Nunes*, Associated Professor,
Iscte - Instituto Universitário de Lisboa

Co-Supervisor:
PhD *Pedro Lopes da Silva Mariano*, Integrated Researcher,
ISTAR-Iscte - Information Sciences, Technologies and Architecture Research Centre

September, 2025

[This page is intentionally left blank.]

Because even during challenging periods,
perseverance and ambition can turn goals into reality.

[This page is intentionally left blank.]

Acknowledgements

I want to express my sincere appreciation to my family and friends, who made this study possible. Special recognition goes to my parents and close family, on whom I could always rely and who always offered comforting words. To my college friends, I extend heartfelt thanks for the long hours and study sessions in the study hall, which contributed both to the completion of our bachelor's degrees and to the intense work carried out during the past months.

A special thank you to Ana, who had the courage to accompany me during our semester abroad and, as always, shared every step of this journey with me, and to the boys, whose company I truly enjoyed. To Afonso and Susana, I sincerely appreciate the time they made available to provide advice and guidance on various matters throughout this study.

To my supervisors, Luís and Pedro, I am deeply grateful for their guidance and crucial support throughout the development of this study.

Last but not least, I would like to thank the Fundação para a Ciência e a Tecnologia, I.P. (FCT), for partially supporting this study through the ISTAR Projects UIDB/04466/2025 and UIDP/04466/2025.

[This page is intentionally left blank.]

Resumo

Estudos relativos ao emprego público português continuam a ser limitados, especialmente no que diz respeito a abordagens para deteção de tendências e antecipação de necessidades futuras. Este estudo aborda esta lacuna através da análise de dados históricos e da avaliação de metodologias de previsão das necessidades de profissionais no setor do ensino público em Portugal.

As evidências produzidas por estes métodos são analisadas em profundidade, não só para avaliar o seu desempenho preditivo, mas também para avaliar a sua viabilidade global e robustez na captação de tendências no sector público. Após a avaliação dos modelos, o estudo fornece projeções do número de posições de ensino futuras, até 2027, comparando o desempenho e relevância de diferentes estratégias de modelização, apoiando assim uma tomada de decisão mais informada e planeamento estratégico no setor.

Os resultados indicam que os modelos de séries temporais superam as abordagens de aprendizagem automática, particularmente quando são incorporadas características externas, como o número de estudantes registados e a taxa de desemprego.

As projeções sugerem uma estabilidade global nos níveis de emprego nos próximos anos, embora também revelem uma tendência pronunciada de envelhecimento no setor. Quando as posições são desagregadas por tipo de contrato legal, espera-se que os contratos por tempo indeterminado sigam uma trajetória estável, enquanto os contratos a termo e os contratos de comissão de serviço, cargo político/mandato apresentam padrões mais voláteis e menos previsíveis.

Palavras Chave: *Previsão de Posições de Trabalho*; *Emprego Público em Portugal*; *Recursos Humanos no setor da Educação*

[This page is intentionally left blank.]

Abstract

Research on Portuguese public employment remains limited, particularly regarding approaches to detect trends and anticipate future needs. This study addresses this gap by analysing historical dynamics and evaluating methodologies for forecasting workforce requirements in Portugal's public education sector. The results produced by these methods are thoroughly analysed, not only to assess their predictive performance but also to evaluate their overall viability and robustness in capturing trends in the public sector.

Furthermore, the study provides projections of teaching positions up to 2027, comparing the performance and suitability of different modelling strategies, thereby supporting more informed and strategic decision-making in workforce planning.

The findings indicate that Time Series models outperform Machine Learning approaches, particularly when external features such as the number of registered students and unemployment rate are incorporated.

Projections suggest overall stability in employment levels in the coming years, though they also reveal a pronounced ageing trend within the workforce. When positions are disaggregated by type of legal contract, permanent contracts are expected to follow a stable trajectory, while fixed-term and commission service or political office/mandate contracts display more volatile and less predictable patterns.

KEYWORDS: *Employment Forecast*; *Portugal Public Employment*; *Education Workforce*

[This page is intentionally left blank.]

Contents

[This page is intentionally left blank.]

List of Figures

[This page is intentionally left blank.]

List of Tables

List of Acronyms

AGR: Annual Growth Rate

AI: Artificial Intelligence

ARIMA: Autoregressive Integrated Moving Average

ARIMAX: Autoregressive Integrated Moving Average with Exogenous Regressors

DGAEP: Directorate-General for Administration and Public Employment

GDP: Gross Domestic Product

GFCF: Gross Fixed Capital Formation

MAE: Mean Absolute Error

MAPE: Mean Absolute Percentage Error

ML: Machine Learning

MLP: Multi-Layer Perceptron

MPE: Mean Percentage Error

MSE: Mean Squared Error

NRMSE: Normalised Root Mean Squared Error

OECD: Organization for Economic Co-operation and Development

RMSE: Root Mean Squared Error

SARIMA: Seasonal Autoregressive Integrated Moving Average

SARIMAX: Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors

SVM: Support Vector Machine

SVR: Support Vector Regression

TS: Time Series

VAR: Vector Autoregressive Model

XGBoost: Extreme Gradient Boost

CHAPTER 1

Introduction

In recent years, there has been a noticeable increase in news headlines displaying the challenges faced by the Portuguese government in managing and addressing public employment issues effectively. The growing number of strikes and interviews with professionals expressing their struggles and lack of understanding has become a recurring topic within the Portuguese community.

According to [1], in the 2022/23 academic year, public education in mainland Portugal served 79% of all registered students, totalling 1,557,785 out of 1,971,312 students. During this period, the public sector needed to enrol 94,645 students in the first grade. A significant challenge faced by this sector is teacher shortages, as there are insufficient educators and professionals overall, with a notable imbalance in their distribution across various regions of the country. This issue is further highlighted in [2], where it is reported that approximately 200,000 students are expected to begin the 2024/25 school year without a teacher for at least one subject, highlighting the severity of the situation. The ageing of the teaching workforce in Portugal has also been noted in international reports. For example, the TALIS 2018 survey [3] reports that 47% of teachers in the country are aged 50 or above, compared to the 34% average across Organization for Economic Co-operation and Development (OECD) countries. This represents a substantial increase from TALIS 2013 [4], where teachers aged 50 or older accounted for 27.9% of the Portuguese workforce, slightly below the OECD average of 30.1%.

Education is the second biggest economic activity in public administrations in Portugal, responsible for 34% of public employment in the fourth quarter of 2024, according to [5]. Kindergarten teachers and primary and secondary school teachers account for 18.8% of total public workers in Portuguese public administrations. These values intensify the need for further exploration of the education sector in Portugal.

In addition to analysing and exploring trends in the available data, there is another question that arises: Is it possible to predict future public employment figures in Portugal based on past records? More specifically, in the education area? The purpose of this research is to comprehend and identify possible solutions to forecast future work positions, establishing whether these methods correspond to a viable solution for the resolution of this problem.

1.1. Motivation

The limited research and investigation concerning Portuguese public employment incites the need to study and explore new approaches to find and analyse any tendencies or orientations within this field. Additionally, this research focuses on addressing the application of Time Series (TS) and Machine Learning (ML) algorithms to make forecasting predictions, evaluating the quality and reliability of such innovative techniques when it comes to anticipating social demands across the public education sector within this context.

1.2. Goals

The objective of this investigation aligns with the intention of finding and analysing patterns in public employment in Portugal, in the education sector, followed by the experiment and evaluation of forecast methods, both based on the analysis of historical data and also through the application of TS and ML methods. Consequently, the results of both approaches will be compared to establish the advantages and disadvantages of using these methods to predict such information, in addition to assessing their forecast capabilities.

1.3. Development

This process is divided into four sequential parts. First, the methods and algorithms to evaluate are defined, building on the conclusions from the Literature Review in Chapter 2 and other relevant considerations. Second, the evaluation metrics for the methods are defined to ensure a concrete and structured assessment process of these algorithms. Next, historical data is used to generate predictions, serving both as a baseline for subsequent forecasting methods and as a benchmark to compare metric values. This step provides insights into the strengths and weaknesses of the following approaches, helping to identify which are better suited to solving the problem.

Subsequently, the selected forecasting approaches, both TS and ML methods, are applied to estimate public employment needs within the context of Portuguese public employment. The results produced by these methods are thoroughly analysed, not only to assess their predictive performance but also to evaluate their overall viability and robustness in forecasting trends in the public sector.

Finally, the approach that demonstrates the best performance in terms of predictive capability is employed to project the evolution of public employment figures for upcoming semesters, thereby supporting more informed and strategic decision-making.

1.4. Scope

This study focuses on analysing trends in the education public employment sector in Portugal, and evaluating the forecasting performance of both TS models and ML algorithms in predicting future workforce needs.

The dataset includes employment records from mainland Portugal as well as the autonomous regions of Madeira and Azores. However, for the forecasting component,

only data from mainland Portugal is used in order to streamline the modelling process. The data range from December 2011 to December 2024 and is drawn from four distinct datasets: one from the Ministry of Education and three others representing separate professional categories: Early Childhood and Primary/Secondary Teachers, Polytechnic Higher Education Teachers and University Teachers. Each category is modelled and forecasted independently, enabling category-specific predictions.

All employment data is sourced from national public records, available through DGAEP - Boletim Estatístico do Emprego Público (BOEP) [6]. In addition, selected external variables are included to enrich the models and improve predictive performance by incorporating complementary explanatory features.

The forecasting methods applied include traditional TS models such as Autoregressive Integrated Moving Average (ARIMA) and Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX), as well as ML algorithms like Multi-Layer Perceptron (MLP) and Extreme Gradient Boost (XGBoost). The study aims to determine how effectively these methods can be used to forecast and plan for future staffing needs in education.

The analysis and model assessment are conducted using *Python* (version 3.11.5) in *Visual Studio Code* and *Google Colab*, with *Microsoft Excel* used for supplementary data processing and results presentation.

It is important to note that this study does not account for qualitative factors such as job satisfaction, motivation or intent to leave the profession. Furthermore, it does not include forecasts for professional categories outside the scope of the education sector, nor does it account for educational institutions that fall outside the public sector.

1.5. Research Questions

In light of the challenges outlined above, it becomes essential to establish a clear research direction for the analysis and forecasting of Portuguese public employment. The following research questions are therefore formulated to guide the investigation process:

RQ1. Can tendencies and patterns be identified in public employment data, in the teaching sector, in the past few years?
The objective of this question is to address any perceived patterns in the various areas of study, allowing an understanding of their meaning and context within the given data and relating these tendencies and patterns to known events.

RQ2. Can future public employment be forecast through the analysis and investigation of related past processes?
The purpose of this question is to understand if there's a way to retrieve information and gain insights into future developments, applying only mathematical and methodical mechanisms.

RQ3. How do TS and ML methods contribute to improving the predictive accuracy in forecasting public employment?
This research question aims to explore the new functionalities that TS and ML methods can bring to this field of study, by evaluating their forecasting results and drawing conclusions about their prognosis capabilities on the subject.

1.6. Organisation of the document

The subsequent chapters are organised as follows: Chapter 2 reports the Literature Review process and its conclusions; Chapter 3 details the datasets employed and provides an exploratory analysis of the public employment data within the education sector; Chapter 4 outlines the methodologies and evaluation metrics employed; Chapter 5 presents the outcomes of the investigation and offers an interpretation of the findings; Chapter 6 summarises the study's conclusions, indicating possible challenges of this study and suggesting directions for future research.

CHAPTER 2

Literature Review

This chapter presents a comprehensive review and critical analysis of existing literature and relevant documents related to the subject of study. It highlights the key findings, identifies strengths and limitations and evaluates their relevance to the current research.

To do this, different databases and directories are used to search for correlated documents, with Scopus[1] and Google Scholar[2] as the sources for the found documents. Several keywords are used to try to detect papers associated with the problem at hand, but only one combination shows effective research results, as the other studies are not relevant to the research problem. The keyword used is '*employment AND forecast*'.

In addition, the combinations '*time series AND forecast*' and '*machine learning AND forecast*' are also employed, not to directly address the research problem, but rather to provide a broader understanding of potential modelling approaches and to assess which forecasting models might be most suitable for addressing the problem under study.

Scopus is a large and trustworthy database for refereed literature. With this tool, about 1,627 records are identified. After a quick examination, 120 records are excluded due to duplication and the remaining 1,507 are brought to a deeper investigation. From this number of documents, only 21 are assessed for eligibility, as the others lack resemblance to the topic under discussion. The remaining records are carefully scanned, and 7 of them are dismissed due to lack of reliability (2), lack of relevance (3) or lack of information (2).

On the other hand, Google Scholar is a web search engine with abundant bibliographic sources. It is important to note that all the documents retrieved from this origin are carefully revised and investigated. There are 21 records found on this platform, as though only 6 are assessed for eligibility. Additionally, one of these is discarded due to a lack of information.

In conclusion, only 19 studies are included in the review and utilised in the context of this study, as illustrated in the PRISMA flow chart [7] presented in Figure 2.1. The three keyword combinations used in the search process are depicted in this figure.

The fact that there are limited documents and research reports on this topic clearly shows the gap in existing research on the subject. Along with the hurdles and limited information, this underscores the difficulty in addressing the issue and its importance in the field.

[1]Abstract and citation database, available at `https://www.scopus.com`, accessed 14 February 2025.
[2]Search engine for scholarly literature, available at `https://scholar.google.com`, accessed 31 January 2025.

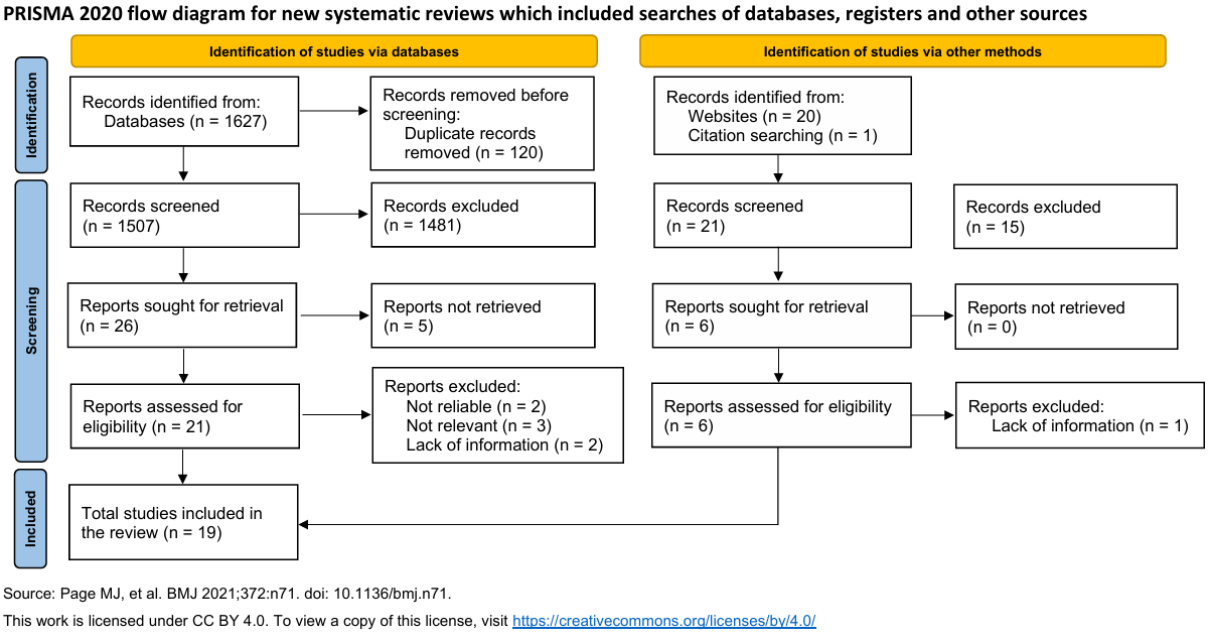


Figure 2.1. PRISMA 2020 Flow Diagram

The following sections address the found papers and their insights into the resolution of the problem under study. Section 2.1 provides a summary of the examined papers; Section 2.2 refers to the studies relating to Portuguese employment, while the following two are divided by the method of resolution of their problem: Section 2.3 describes Historical Analysis approaches; Section 2.4 TS and ML applications.

2.1. Brief summary of retrieved records

For analysis purposes, the retrieved records are examined from a quantitative perspective. Figure 2.2 presents the number of studies published per Year of Publication that met the inclusion criteria. The distribution shows that early research activity in this domain is sparse and intermittent throughout the 1990s and early 2000s, followed by a gradual increase in publications after 2005.

Figure 2.3 illustrates the relative word frequency in the reviewed studies, to understand the importance of each feature or definition in the context of this study. Terms such as *model*, *data*, *forecasting* and *time series* dominate the corpus, confirming the strong orientation of the field.

2.2. Portugal employment-related records

In the documents retrieved, a small portion has the Portuguese public and private employment as their subject environment in their investigation. These studies aim to analyse trends in the field, along with predicting the necessities and requisites of the designated sectors in the future.

The study in [8] does not aim to produce forecasts, but rather to identify trends in construction contracts within Portuguese public employment. The analysis is based on a descriptive and systematic exploration of 5,172 public procurement contracts spanning the

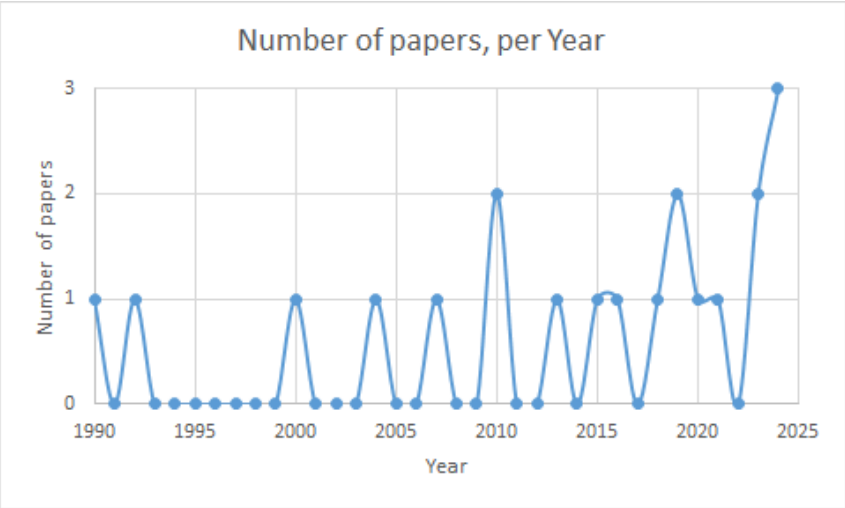


Figure 2.2. Number of records retrieved, per Year of Publication

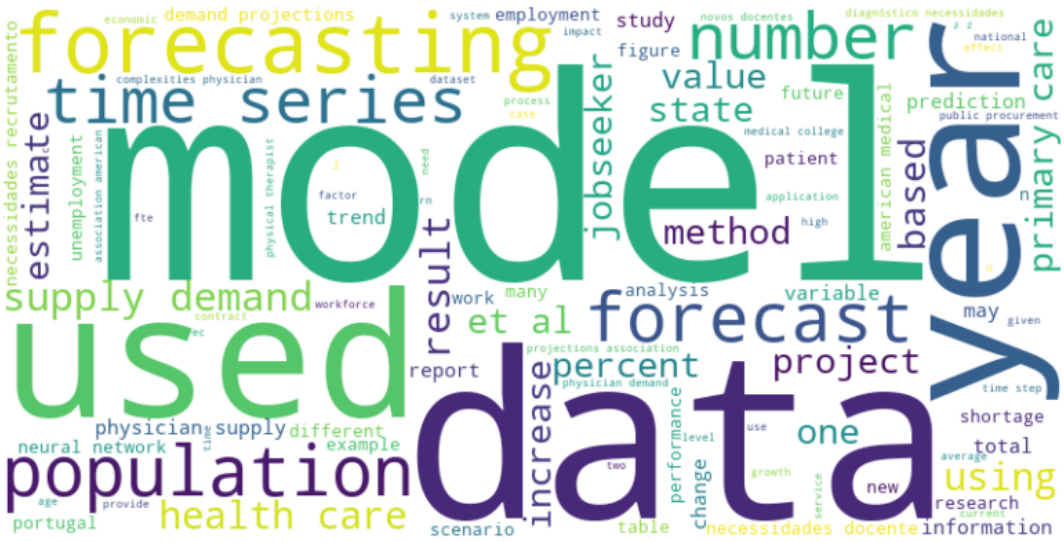


Figure 2.3. Word Frequency in studied records

period from 2015 to 2022. The dataset is structured into several interrelated subchapters, which are examined within their respective sections. Two interpretative approaches are then applied to extract deeper insights: one focusing exclusively on the primary feature, the contract price, and another employing a multifactorial criterion to enable a broader evaluation. The results highlight current tendencies in public procurement contracts, identifying major patterns and proposing best practices for structuring selection criteria according to the type of project. Overall, the performance of construction projects in Portugal during the studied time frame is assessed as positive. However, the findings indicate that the award criteria are generally not correlated with the final project price, and the use of multi-factor assessment criteria does not necessarily lead to better performance compared to contracts awarded solely based on price.

The other papers reviewed in this section have as an objective the prediction of future employability. Two of these approaches rely exclusively on historical data to forecast

their respective study problems, whereas their dilemmas rely on the health and education sectors, respectively. In both of these research papers, the data from either of these is thoroughly analysed, followed by an intense exploration and improvement of a model to assess the forecasting necessities.

The authors in [9] implement a ML model in order to predict construction project compliance. Due to an imbalance in the target feature, three different techniques are applied to balance the dataset: Oversampling, Undersampling and the Synthetic Minority Oversampling Technique (SMOTE) method, with an Adam Artificial Neural Network (ANN) algorithm benchmark for evaluation purposes. Next, all data treatment techniques are executed, and a model is defined using a feature selection process. Five algorithms are tested to achieve better results: Adam ANN, Random Forest, Support Vector Machine (SVM), XGBoost and K-Nearest Neighbours (KNN). The results show that the SMOTE method provides better outcomes when balancing the dataset. Using this tool, the algorithm Adam ANN outperformed others, with a precision rate of 68.1%. While the study reveals that ML algorithms cannot accurately predict budget compliance using employment data, they can still offer project owners valuable insights into the most relevant criteria, supporting informed decision-making.

According to [10], in 2022, Portugal had 573.5 physicians licensed to practice per 100,000 inhabitants, ranking second in the European Union. The authors in [11] try to predict the number of needed physicians in the following 12 years, with a model focused on the number of placed students' trends in the medicine integrated master's in the country, along with retirement and death rates. Furthermore, employment leaves and career entrances due to immigration are also considered. These projections estimate that, by 2034, there will be a ratio of 602 physicians per 100,000 inhabitants, compared with the 502 physicians per 100,000 inhabitants in 2024.

In relation to the education sector, [12] creates a forecast model based on the supply and demand of teacher resources. The necessity for faculty members is calculated in regard to the number of children and young people registered in each school year, along with birth rates. Additionally, the transition rates by school year and the percentage of students that transfer from public to private schools and vice versa are also considered. The supply of teachers remaining in the profession is then subtracted from this projection. The predictions show that the number of students registered in public schools should substantially diminish in the next years to 960,919 students in 2030/31, a drop of 15% compared to 2018/19. A substantial reduction in teaching staff is also expected in the coming years due to scheduled reductions and retirements. It will be necessary to recruit an average of 3,450 new teachers per year (a total of 34,508 until the school year of 2033/34) to address this shortage, with an emphasis on pre-school education.

2.3. Forecasting based on Historical Analysis Approach

A traditional approach to forecasting future data relies on the analysis of historical records related to the subject under study. This process involves the application of methodological

and mathematical techniques to derive solutions for the problems being addressed. The following subsections present different techniques commonly employed to forecast future values using this approach.

2.3.1. Previous Approaches

In the context of employment data analysis, [13] implements a model to analyse dataset characteristics and predict whether employees in Taiwan would choose to stay or leave a company. The authors employ a robust methodology that combines both qualitative and quantitative approaches. Logit and Probit models are tested, using Maximum Likelihood Estimation to determine the free parameters.

In the training dataset, the Logit model outperformed the Probit model, achieving a success rate accuracy of 74.0%, compared to 72.0% for the Probit model. However, in the testing dataset, both models reached a success rate accuracy of 71.9% for predicting employee turnover, although the Logit model had a lower Queries per Second (QPS) value than the Probit model, proving its effectiveness in understanding employees' intentions to leave a company or organisation in such a context.

When predicting future workforce needs, various approaches have been employed for effective assessment. [14] uses demographic projections and labour force analysis to evaluate the anticipated workforce requirements in Europe. In a different context, [15] implements a model based on age distribution and total supply for registered nurses, to project their future age distribution and total supply up to 2020, utilising data from the previous 25 years.

The author in [14] begins by analysing birth rates, the ageing population and overall population decline. The study continues with an evaluation of past and future trends, drawing projections from the United Nations and Eurostat. The potential workforce is determined by the number of people entering and exiting the labour market, while the actual workforce comprises those who are economically active. In the European Union, the available workforce started at 145 million in 1990 (excluding East Germany) and peaked at 146.9 million in 2000, reflecting a 1.3% increase. After this peak, the workforce gradually began to decline until 2010, with a more rapid decrease projected until 2025. Notably, no reduction in the total labour force is evident in the specific case of Portugal.

In [15], an analysis of the observed changes in the size and age of the nursing workforce over time is conducted by breaking them down into three distinct components: population, cohort and age effects. This investigation uses variance analysis to estimate the parameters of the prediction equation, evaluated with an adjusted Coefficient of Determination (R^2) value of 0.82. The findings suggest that the coming years would see steady growth, with the overall number of full-time nurses per capita reaching a peak in 2007, followed by a decline leading to 2020. By that year, it is projected that over 40% of registered nurses will be above the age of 50.

It is essential to note that these studies date back to the 1990s, highlighting the differences between the models employed then and those used today. More recent assessments

typically involve calculating supply and demand, particularly to predict potential workforce shortages in the future.

2.3.2. Supply and Demand Approach

Both [16] and [17] implement a supply-demand model to project future workforce needs, using data from the United States.

In [16], the supply and demand of physical therapists are defined primarily through age-population projections, population growth and historical employment rates. A linear regression analysis is conducted using these definitions, resulting in a model evaluated with an R^2 metric of 0.998. The results suggest that shortages in physical therapists are expected to increase over the next two decades.

In contrast, [17] determines supply projections by summing specific personnel characteristics, including the current number of physical therapists and the number of new graduates, while subtracting the number of graduates who did not pass the licensure examination. Demand is characterised by the projected U.S. population with healthcare insurance, multiplied by a demand ratio for any given year. A selected number of variables are chosen, and the model is developed using the *STELLA* software, available in [18].

These studies confirm the existence of physical therapist shortages in many states, with projections indicating that these shortages are expected to increase by the year 2030. By that time, 48 states are anticipated to exhibit physical therapist shortage ratios with a standard deviation value greater than 2.99.

More recently, [19] also implements a supply-demand model to forecast the physician workforce in the United States. This study, conducted in the aftermath of the COVID-19 pandemic, provided a new perspective on the future of this sector. The model is based on the characteristics of the current physician workforce, the annual number of newly trained physicians, hours worked and retirement patterns. It also accounted for COVID-related deaths during the first year of the pandemic, as well as trends in decreased immigration and declining birth rates.

To further explore the issue, two versions of this scenario are modelled: one with a 1% annual growth in the number of new physicians entering the workforce, and one without. Additionally, the scenarios consider physicians retiring two years earlier or two years later than they currently do, due to factors such as financial issues, health concerns and burnout. The supply-demand comparison generates 48 sets of projections for future supply adequacy, taking into account various specialities and overall physician availability. To ensure a more reliable range of projections, the highest and lowest quartiles of supply adequacy projections are excluded.

The findings indicate a shortage of physicians over the next 10 to 15 years, with a projected shortfall of between 13,500 and 86,000 physicians by 2036, which includes a shortage of 20,200 to 40,400 primary care physicians. Furthermore, it is expected that COVID-19 will transition from a pandemic to an endemic state, with an estimated 110

million to 220 million COVID-19 cases annually in the future, further increasing the demand for physicians.

2.4. Forecasting based on Time Series and Machine Learning Algorithms

A more innovative approach to estimating future employment involves two distinct methodological paths: the application of TS forecasting techniques and the use of ML algorithms. TS methods rely on historical data patterns to project future trends, while ML and Artificial Intelligence (AI) techniques leverage complex data structures and learning mechanisms to enhance predictive capabilities. Both approaches are explored in the research papers described in the following subsections.

2.4.1. Time Series Forecasting

A TS is a sequence of data points collected at equally spaced intervals over time. Subsequently, TS forecasting is the process of analysing these data using statistical methods and models to make predictions and support strategic decision-making.

The authors in [20] present forecasting results based on TS applications in Denmark, by conducting a competition among various econometric and TS models to determine the most accurate forecasting method for international tourist expenditure in the country. This research tests six econometric models, including Vector Autoregressive Model (VAR), Autoregressive Distributed Lag Model (ADLM), ARIMA for benchmarking and Error Correction Model (ECM). The models are trained using data from 1969 to 1993 and validated with data from 1994 to 1999, alongside performance evaluations using Mean Absolute Percentage Error (MAPE) and Root Mean Squared Percentage Error (RMSPE) metrics. Upon selecting the VAR model as the best predictor for longer-term forecasts of tourism expenditure, the study employs an input-output model to estimate the impact on employment. The results indicate an expected growth of 4.0% in total foreign tourist expenditure over the next five years and a further 3.2% in the subsequent five years, primarily impacting employment in retail, hotels and restaurants.

Additionally, another study also proposes an export-based ECM as a TS approach, available in [21]. This model, along with others, incorporates long-term relationships between export-based and local employment to enhance forecasting accuracy. The authors train all models using monthly employment data from eight North American cities and assess them through out-of-sample forecasts. The performance of these models is measured using Root Mean Squared Error (RMSE) and the Standard Deviation of Forecast Error. The Dickey-Fuller and Augmented Dickey-Fuller (ADF) tests indicate that export and local employment are co-integrated in most cities, with the ECM model significantly improving employment forecasting by incorporating long-term economic relationships.

2.4.2. Statistical and Machine Learning Based Forecasting

On another approach, [22], [23] and [24] implement models with a focus on the identification of patterns in employment data and statistical profiling. This feature refers to the use of mathematical methods to analyse historical data and understand the sequences,

distributions and trends in employment activities, helping the prediction of future demand and employment needs more accurately.

The study by [22] aims to differentiate job seekers who are likely to become long-term unemployed from those who are likely to find work quickly. Each country adopts a different model to address this issue, using statistical methods to predict labour market disadvantages based on socio-economic characteristics such as age and gender, as well as three employment barriers: motivation, job capability and opportunities.

In Flanders, Belgium, a random forest model is employed to estimate the probability of long-term unemployment. This model is designed to be flexible, allowing it to be updated regularly to maintain accuracy, achieving an accuracy rate exceeding 67% and an Area Under the Curve (AUC) value of approximately 0.76. With the same goal, Denmark and New Zealand implement predictive models. The Danish Agency for Labour Market and Recruitment (STAR) develops a profiling model that utilises ML, specifically decision tree classification, to forecast the likelihood of long-term unemployment. This model boasts an accuracy of over 60%. Meanwhile, New Zealand employs Random Forest and Gradient Boosting techniques to predict lifetime income support costs and assess the effectiveness of case management services, achieving AUC values ranging from 0.63 to 0.83.

The study by [23] focuses on detecting suspicious one-bid tenders using text mining and ML methods, following the CRISP-DM framework. Text-mining techniques are applied to extract relevant parts of the documents, followed by text processing using Natural Language Processing. While it is not possible to predict if a tender is corrupt, the authors propose models to determine if a tender has only one bid, which may signal suspicion. Three common text classification algorithms are tested: Naïve Bayes, Logistic Regression and SVM, with performance metrics including Accuracy, Precision, Recall and Area Under the Receiver Operating Characteristic (AUC-ROC) curves.

Logistic Regression emerges as the best-performing algorithm overall, achieving the highest accuracy and AUC score. To further enhance accuracy, the dataset is categorised into eight employment categories, with the Information Technology Services and Health sectors demonstrating the highest accuracy and recall rates. In predicting outcomes related to health and social work services, both SVM and Logistic Regression perform better, while in most other cases, the Naïve Bayes algorithm yields better results.

Finally, [24] propose a feature selection prediction model based on an improved bat algorithm combined with a SVM to predict slow employment. Slow employment refers to the phenomenon of college graduates choosing not to seek employment or further education immediately after graduation. The bat algorithm is an optimisation method inspired by the echolocation behaviour of bats. This algorithm is enhanced by a Gaussian distribution to improve search capabilities and incorporates an elimination strategy to enhance population diversity and avoid local optima. This improved version is referred to as 'GEBA', which is further modified into 'bGEBA' to facilitate discrete feature selection.

The algorithm identifies the optimal feature subset, which is subsequently used to train and predict outcomes with the SVM classifier. The results demonstrate that the bGEBA-SVM model achieves the highest prediction performance, with an accuracy of 93.86% and a F-measure of 93.36%, surpassing traditional models and providing insights into the most significant factors influencing graduate employment.

2.4.3. Forecasting using Time Series and Machine Learning models

Several studies have been performed using both TS and ML methods, enabling a possible comparison between the models' performance and capabilities.

The author in [25] implements a model to predict the monthly number of individuals registered with the Swedish Public Employment Service who will successfully obtain employment. This is done using both the Seasonal Autoregressive Integrated Moving Average (SARIMA) model and the Long Short-Term Memory (LSTM) model, also exploring the possibility of utilising multiple TS simultaneously during certain experiments, always including the TS intended to forecast. The TS with the highest correlation values is selected using Pearson's correlation for these experiments.

After preprocessing and feature selection, the LSTM model is defined while experimenting with various model parameters to identify the optimal representation. The best multivariate LSTM model demonstrates a 47.38% improvement over the persistence model, a straightforward baseline approach that assumes the employment numbers for the next month will remain the same as those of previous months. Additionally, it shows a 7.05% improvement over the univariate LSTM model, confirming its superior performance.

In [26], a comprehensive review of TS forecasting models for industrial applications is presented. The study discusses classical statistical models, such as ARIMA, SARIMA, Exponential Smoothing and Bayesian Structural Time Series (BSTS), which are valued for their simplicity and interpretability but often struggle with complex, non-linear or high-dimensional data. ML models, including SVM, Regression Trees, Random Forests and KNN, and Deep Learning (DL) models like ANN, RNN and LSTM, provide greater flexibility and can capture non-linear patterns and long-term dependencies, though they require larger datasets and careful tuning. Hybrid models that combine statistical and ML/DL approaches are highlighted as particularly promising, leveraging the strengths of each paradigm.

The review emphasises that no single model performs optimally in all contexts. Therefore, model selection depends on the data characteristics, dimensionality and forecasting objectives. Overall, statistical models remain competitive for simpler or smaller datasets, while ML and hybrid models are advantageous in complex, multivariate or highly non-linear scenarios.

In the study conducted by [27], a ML-based prediction system is developed to forecast the number of patients affected by COVID-19. Four standard forecasting models are employed: Linear Regression (LR), Least Absolute Shrinkage and Selection Operator

(LASSO), SVM and Exponential Smoothing. Each model generates three types of predictions, namely the number of newly infected cases, the number of deaths and the number of recoveries over the subsequent ten days.

The results indicate that Exponential Smoothing outperformed all other models, followed by LR and LASSO, which demonstrates good predictive accuracy across all types of predictions. In contrast, SVM exhibited poor performance in the prediction scenarios, likely due to the instability of dataset values. These findings highlight the potential of combining statistical and ML approaches for short-term epidemic forecasting, particularly when data is limited.

2.4.4. Forecasting using a data mining framework

The authors in [28] present a data mining framework that utilises search engine query data to predict unemployment rates. Within this framework, a set of data mining tools, including neural networks and Support Vector Regression (SVR)s, is developed to forecast unemployment trends.

After extracting search engine query data related to employment activities, a feature selection model is suggested to reduce the dimension of the query data, using the top 100 most correlated features. Several neural networks and SVRs are employed to model the relationship between unemployment rate data and query data, and a genetic algorithm is used to optimise the parameters and refine the features simultaneously.

Finally, the selective predictor with the best feature subset and proper parameters, defined by a cross-validation method, is used to forecast the unemployment trend. According to the metrics results analyses, the v-SVR with Radial basis function kernel in iteration 3 model is chosen as the model for the final prediction, with an RMSE value of 68,182.55, an MAE value of 54,241.10 and a MAPE value of 12.54.

Although various literature reviews and studies have been conducted on related topics, none correspond closely to the specific objectives of this research. This reveals a clear gap in the existing literature and reinforces the need for further investigation into this particular area.

CHAPTER 3

Data Understanding

The primary data source for this study is the Public Employment Statistics Bulletin from June 2025, published by the Directorate-General for Administration and Public Employment (DGAEP). The data is publicly available through *Boletim Estatístico do Emprego Público (BOEP)* [6].

3.1. Dataset Description

Data is extracted from the first ('Public Administrations') and third ('Other indicators') worksheets of the bulletin. The first sheet provides employment data from various Portuguese ministries, including detailed figures for different educational roles, while the third sheet includes regional employment indicators.

Within each sheet, the data corresponding to the education sector is extracted. In the first case, the data from the scholarship ministries (Section Q.1.1 of the source) is retrieved, along with the data from the different teaching positions (Section Q.1.3 of the source), defined as follows:

- *Educadores de Infância e Docentes do Ensino Básico e Secundário* (Early Childhood Educators and Primary/Secondary Teachers),
- *Docentes do Ensino Superior Politécnico* (Polytechnic Higher Education Teachers),
- *Docentes do Ensino Universitário* (University Teachers).

Each worksheet contains data on the number of work positions, disaggregated by ministry or teaching position, gender and age group, recorded for each semester. Other indicators, such as the *Youth Index*, are also included.

In the second dataset, the information extracted refers to the geographical distribution of public education and teaching employment across NUTS I, II and III regions (Section Q.3.2 of the source), along with supplementary indicators such as the number of school establishments by region. NUTS is a hierarchical system used by the European Union to divide countries into statistical regions, facilitating the collection, analysis and publication of regional statistical data.

Both datasets span the period from December 2011 to December 2024 and are consistently broken down by age group and gender, reporting the number of work positions in each demographic segment for every semester.

Supplementary statistics are incorporated to support the analysis of the source datasets, namely the number of residents in Portugal, disaggregated by NUTS III region and year.

These figures allow for a more meaningful interpretation of work positions and other variables, both in absolute terms and relative to the population, for example, standardised per 1,000 inhabitants. This data is extracted from the Portuguese National Institute of Statistics [29].

3.1.1. Workforce Data by Type of Contract

Another dataset is employed in this study, as it contains the number of work positions by Type of Contract in Portugal, specifically for Early Childhood Educators and Primary/Secondary Teachers working in public education establishments. This dataset is obtained directly from DGAEP upon request.

The dataset covers the period from December 2011 to March 2025 and is structured by NUTS regions and municipalities, trimester and type of contract, distinguishing the following categories:

- *Contrato a termo* (Fixed-Term Contracts),
- *Contrato por tempo indeterminado* (Permanent Contracts),
- *Comissão de serviço, cargo político/mandato* (Commission Service or Political Office/Mandate Contracts).

For consistency with the other datasets employed in this study, the analysis is conducted at the NUTS III regional level, encompassing all 26 NUTS III regions of Portugal. Additionally, only data from the second and fourth trimesters (June and December) is considered.

3.1.2. Feature Integration

To enhance the performance of the forecasting process, a range of additional explanatory variables is incorporated into the models. These features aim to reflect relevant political and socio-economic dynamics within Portuguese society, which may impact the target variable, the number of positions in the public education workforce.

The selected features include Gross Domestic Product (GDP), Gross Fixed Capital Formation (GFCF), Public Debt, Unemployment Rate, Minimum Wage and the total number of Registered Students in Portugal.

The forecasting process takes advantage of both historical and estimated data. Most of the projected data is obtained from *Banco de Portugal*'s biannual economic bulletins, covering the period from 2011 to 2025. However, for the variables Public Debt and Unemployment Rate, projections are obtained entirely from the OECD's biannual economic bulletins, as *Banco de Portugal* did not provide full coverage over the required time span.

The projected data used in this study originates from a series of biannual publications by accredited institutions, which provide medium-term forecasts for selected socio-economic variables. These projections typically extend up to 2026 or 2027, depending on the institution. In parallel, the historical data consists of a list for each variable, containing the actual observed values for each country. The projected data is available through [30], while the actual historical data is extracted from the Fundação Francisco Manuel

dos Santos (FFMS) statistical portal [31], accessed on 19 June 2025. The names of the specific datasets used include:

- GDP - *Taxa de crescimento real do PIB (%)*
- GFCF - *Investimento público, em % do PIB*
- Public Debt - *Dívida pública, em % do PIB*
- Unemployment Rate - *Taxa de desemprego por sexo, grupo etário e nacionalidade (%)*
- Minimum Wage - *Salário mínimo nacional*
- Number of Registered Students for Early Childhood Educators and Primary/Secondary Teachers - *Alunos matriculados do pré-escolar ao secundário por sexo, subsistema de ensino e nível de ensino*
- Number of Registered Students for Polytechnic Higher Education Teachers and University Teachers - *Alunos inscritos no ensino superior por subsistema de ensino, tipo de ensino e ciclo de estudos*

3.2. Exploratory Data Analysis

In order to better understand the source data, an exploratory data analysis is conducted. The analysis begins with the complete dataset from the Portuguese ministries that includes the education sector, followed by the analysis of the data across the different teaching positions, and finally by region. The goal of this analysis is to identify trends, patterns and potential relationships within the data by examining its components and exploring correlations between variables.

3.2.1. Ministries Dataset

The data from the Portuguese education ministries, including the regional administrations of Madeira and Azores, is structured by administration, ministry, age group, gender and semester, comprising a total of 972 records. Each record corresponds to a unique combination of these features. It is relevant to mention that each administration is linked to a specific education ministry, as each region has its own department.

The analysis of the relationship between the features and the target variable, the number of work positions, shows that age group, administration and ministry exhibit the strongest relationships with the target, as indicated by higher F-scores from Analysis of Variance (ANOVA), as described in Figure 3.1. Gender and semester exhibit a weaker association with the number of work positions, indicating a limited impact on the target variable. Since the semester is a numerical feature, defined by the combination of year and semester, it is transformed into a categorical variable to better assess its relationship with the target feature.

The distribution of work positions by ministries highlights the disparity in workforce size between mainland Portugal and the autonomous regions of Madeira and Azores, as illustrated in Figure 3.2.

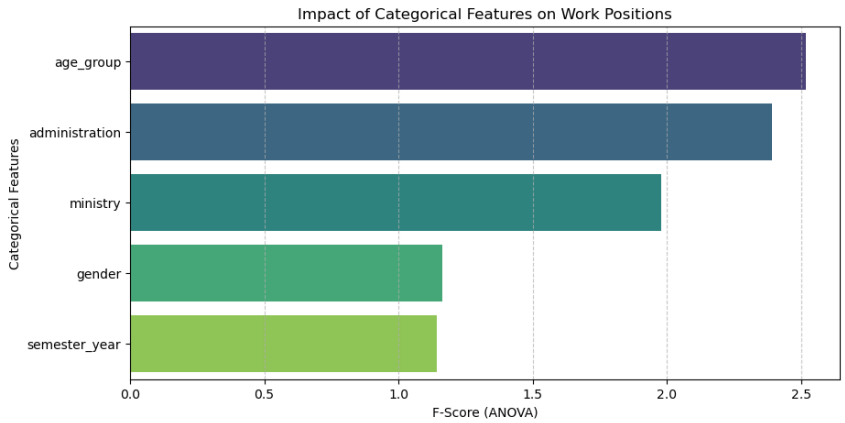


Figure 3.1. Relationship between Target variable and other columns, in the dataset per Ministry

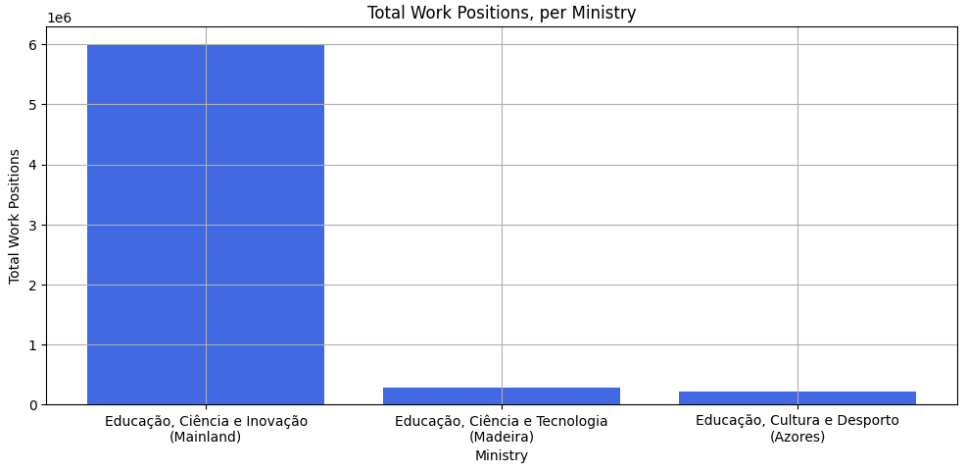


Figure 3.2. Distribution of Work Positions, per Education Ministry, in Continental and Insular Portugal

The data retrieved reveals a slight descent in the total number of work positions in education ministries, from 2011 to 2015. After 2015, the value has been increasing until the end of available records, the second semester of 2024, reaching the maximum value of 258,596 work positions, as described in Figure 3.3.

This reduction can be explained by the severe economic and social crisis Portugal experienced during the Troika's intervention (a political regime led by the European Commission, the European Central Bank and the International Monetary Fund). During this regime, teacher recruitment is drastically cut and Fixed-Term Contracts are not usually renewed, as over 31,000 teachers left the profession between 2011 and 2014 [32]. In addition, the sector is affected by early retirements as part of austerity policies. Faced with fewer job opportunities and significant salary cuts, many teachers opted to emigrate to other countries in search of better working conditions.

The distribution of working positions per gender and semester highlights the predominance of women in the workforce within the education sector ministries. Variations in the

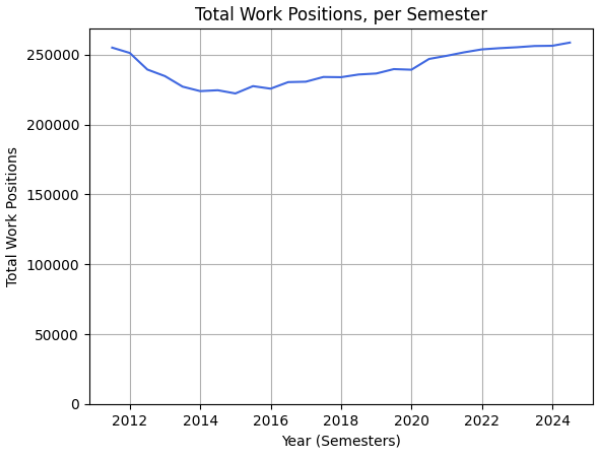


Figure 3.3. Total Education Ministries Work Positions, per Semester

total workforce over time are mainly driven by oscillations in the number of women employees, while the number of men has remained relatively stable, as illustrated in Figure 3.4.

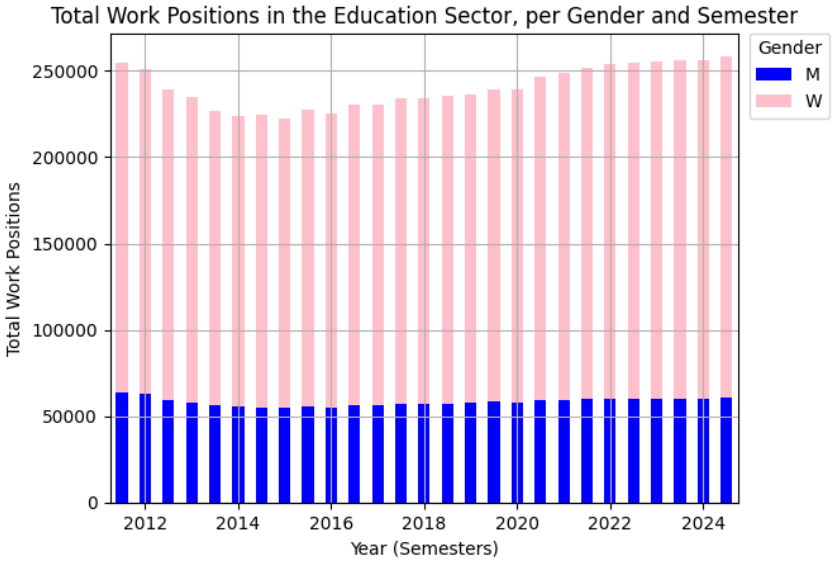


Figure 3.4. Distribution of Work Positions, per Gender and Semester

When analysing the number of work positions per age group, as presented in Figure 3.5, there is a clear predominance of the intermediate age groups when compared to the younger (under 24 and 25–34) and older (over 65) categories. The group under 24 years consistently shows the lowest values, with a mean of approximately 900 workers per semester. In contrast, the other age groups register significantly higher values, with averages in the thousands or even tens of thousands. The oldest age group, in particular, exhibits an ascending pattern, which reinforces the idea that the teaching workforce is ageing significantly by semester.

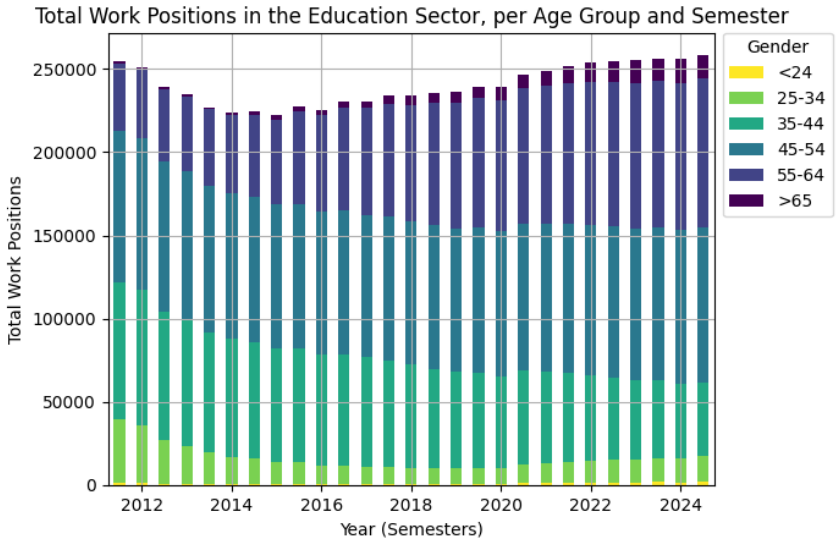


Figure 3.5. Distribution of Work Positions, per Age Group and Semester

3.2.2. Datasets by Type of Teaching Position

When splitting the dataset by type of teaching position, Early Childhood Educators and Primary/Secondary Teachers, Polytechnic Higher Education Teachers and University Teachers, the analysis remains consistent with the ministries dataset.

The relationship analysis results indicate that the type of position holds the strongest association with the number of work positions, as illustrated in Figure 3.6, while other features show limited relevance to the target variable.

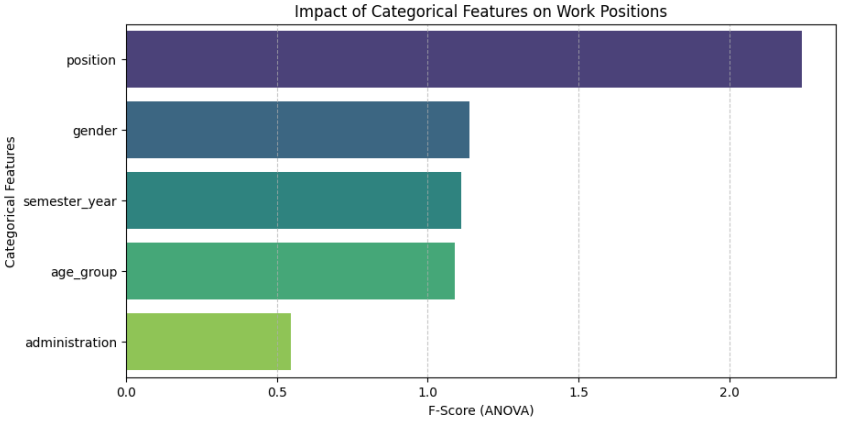


Figure 3.6. Relationship between Target variable and other columns, in the dataset per Type of Position

The distribution of work positions by type of position and semester highlights the significant discrepancy between the number of positions for Early Childhood Educators and Primary/Secondary Teachers, with a mean of 137,015, compared to University Teachers, with a mean of 15,129, and Polytechnic Higher Education Teachers, with a mean of 10,050, as shown in Figure 3.7. This difference is supported by the greater number of

kindergartens and schools compared to universities and polytechnic institutions in Portugal.

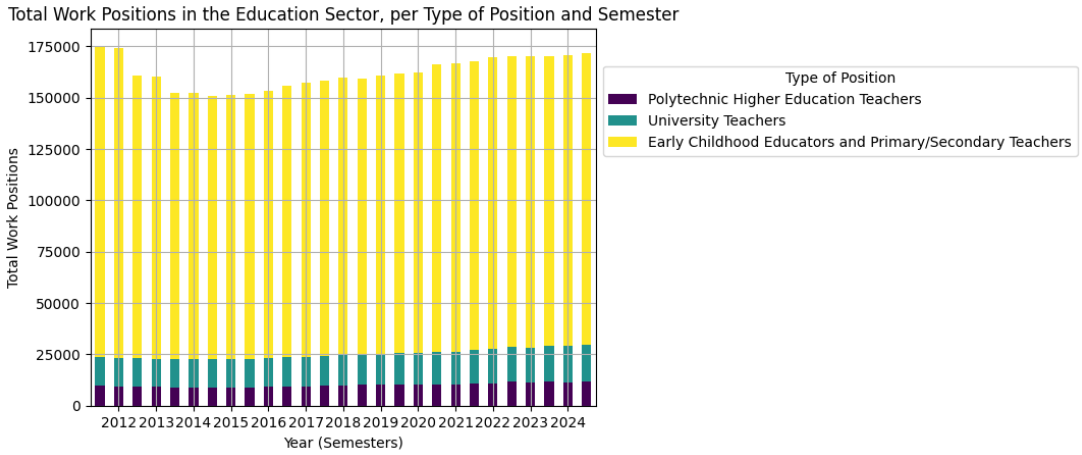


Figure 3.7. Distribution of Work Positions, per Type of Position and Semester

The data corresponding exclusively to mainland Portugal exhibits the same overall trends as previously described. The distributions and patterns across semesters and categories follow similar curves to those observed in the complete dataset, reinforcing the consistency of these tendencies.

3.2.3. Data from different Portuguese Regions

This dataset is used exclusively for analytical purposes, providing an overview of the education workforce and valuable insights into their characteristics, and is not applied in any forecasting exercises. The forecasting analysis is instead conducted with contract-specific datasets at the NUTS III regional level, as described in Section 3.1.1, with the corresponding analysis presented in Section 3.2.4. Although the datasets differ, in some regions, particularly the smaller ones, the number of work positions and the total number of contracts align, producing equivalent results and thus complementing each other. In contrast, larger regions tend to show higher values in the contracts dataset, compared to the dataset by regions.

The data by regions is described based on the NUTS III classification, exclusively for the position of Early Childhood Educators and Primary/Secondary Teachers. Since the population statistics for Portugal aggregate the NUTS regions of *Grande Lisboa* and *Península de Setúbal*, these are represented in the plot as a single region, referred to as *Área Metropolitana de Lisboa*. Furthermore, the available data only extends up to 2023, so the presented plots likewise cover this period.

Analysis shows that, as expected, the regions with the highest numbers of kindergarten teachers and primary and secondary school teachers per 1,000 registered students are those with larger populations, specifically in the metropolitan areas of Lisbon and Porto. This pattern remains consistent throughout the period under study, although some declines between 2011 and 2015 are noticed, as illustrated in Figure 3.8.

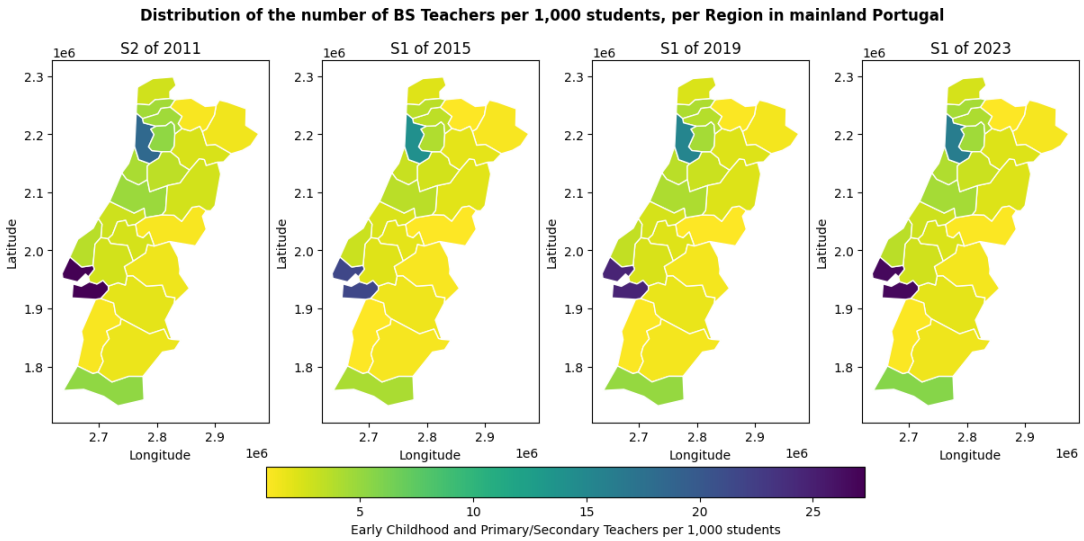


Figure 3.8. Evolution of Early Childhood Educators and Primary/Secondary Teachers, per NUTS III Regions, per **1,000 students**

In contrast, when considering the number of these teachers per 1,000 inhabitants, the mentioned regions rank among the lowest. From the second semester of 2011 to the first semester of 2015, the number of such teachers declined nationwide. After 2015, all regions experienced growth, with particular emphasis on *Alto Alentejo* and *Baixo Alentejo*, which report the highest number of teachers per 1,000 inhabitants in 2023, as illustrated in Figure 3.9.

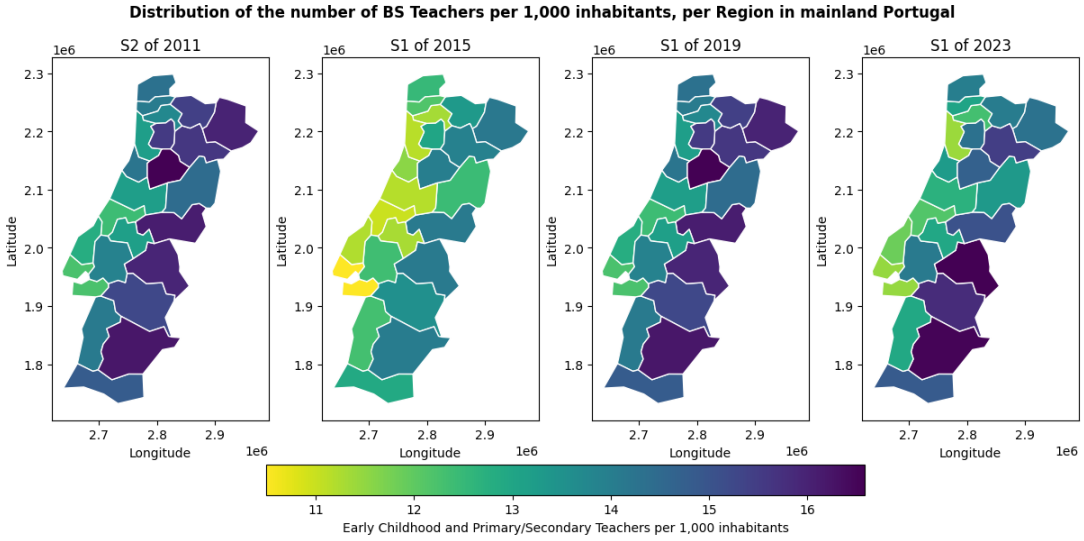


Figure 3.9. Evolution of Early Childhood Educators and Primary/Secondary Teachers, per NUTS III Regions, per **1,000 inhabitants**

3.2.4. Workforce Data by Type of Contract

The dataset used to forecast the number of contracts for Early Childhood Educators and Primary/Secondary Teachers provides the distribution of work positions in this sector,

broken down by NUTS III Region and Legal Relationship (referred to as Type of Contract in the analysis). Consequently, the following analysis focuses exclusively on this specific position. The target variable is the number of *Public Administration Managers and Employees*, which, for simplification purposes during analysis and forecasting, is referred to as the 'Number of Contracts'.

The temporal distribution of the number of contracts is illustrated in Figure 3.10, as Permanent Contracts clearly constitute the majority, accounting for an average of 82.16% of all contracts, followed by Fixed-Term Contracts at 17.42%, and Commission Service or Political Office/Mandate Contracts, which represent only 0.42% of the total.

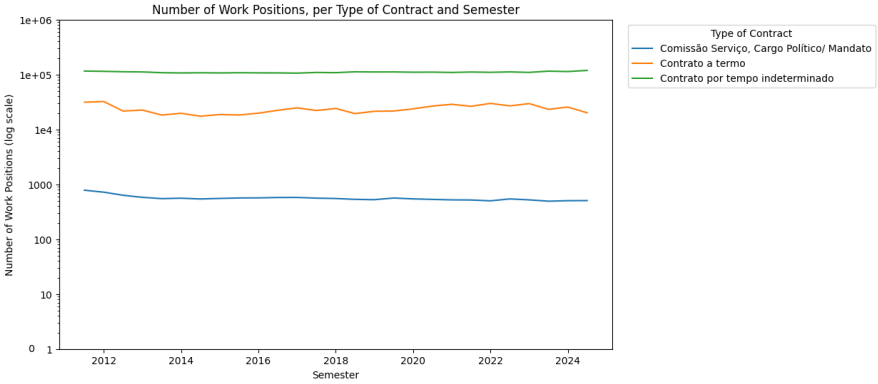


Figure 3.10. Number of Work Positions, per Type of Contract and Semester

Figure 3.11 displays the average number of contracts for each region, highlighting the clear predominance of work positions in the largest regions of the country, in terms of population. In particular, *Grande Lisboa*, the capital, accounts for an average of 16.06% of all contracts, followed by *Área Metropolitana do Porto*, the second largest city, with 14.53%. At the lower end, *Alentejo Litoral* represents 0.91%, while *Alto Tâmega e Barroso* and *Beira Baixa* each account for 0.90% of the total number of contracts.

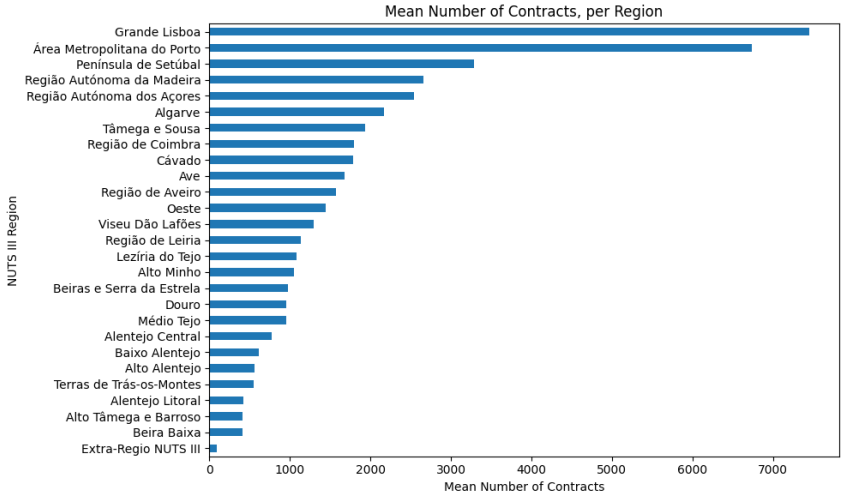


Figure 3.11. Mean Number of Contracts, per Region

The following illustrations portray the evolution of the number of work positions by each Type of Contract, between the years of 2012 and 2024. For Permanent Contracts, in Figure 3.12, the values have remained stable throughout the period under study, showing no significant fluctuations, which indicates a consistent pattern for this type of contract. *Grande Lisboa* and *Área Metropolitana do Porto* stand out as having the highest number of Permanent Contracts in the country, with a substantial margin compared to other regions, followed by *Península de Setúbal* in the south of Lisbon, in third place.

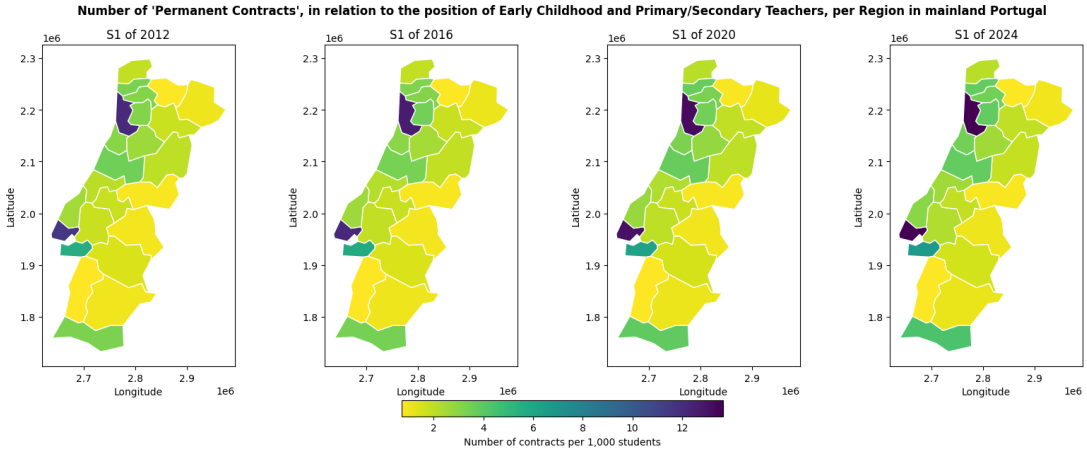


Figure 3.12. Evolution of **Permanent Contracts** per NUTS III Regions, per 1,000 students

In the context of Fixed-Term Contracts, described in Figure 3.13, as with Permanent Contracts, *Grande Lisboa* and *Área Metropolitana do Porto* also stand out as the regions with the highest number of contracts. These regions experienced a slight decline between 2012 and 2016, followed by growth after this period. A similar pattern can be observed in other regions, such as *Tâmega e Sousa*, located to the east of *Área Metropolitana do Porto*, Algarve, in the southern end of the country, and *Península de Setúbal*, directly below *Grande Lisboa*.

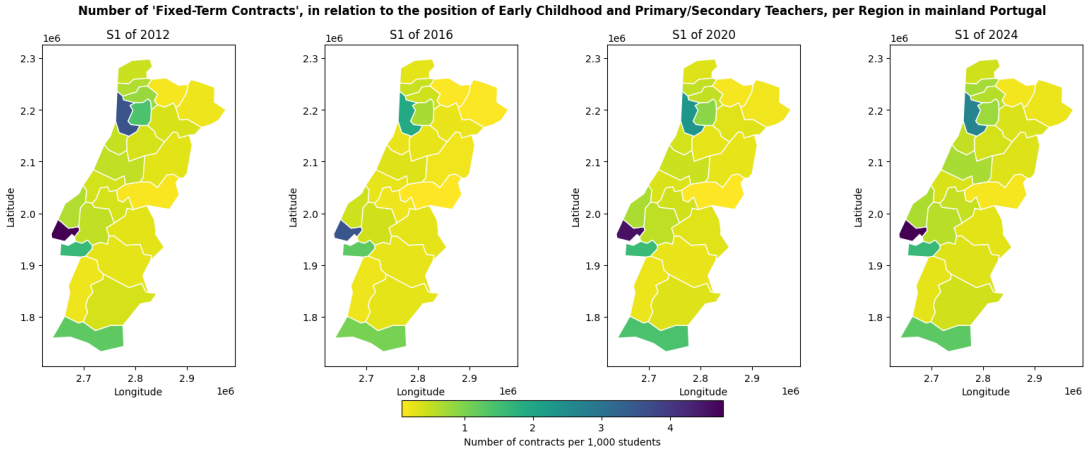


Figure 3.13. Evolution of **Fixed-Term Contracts** per NUTS III Regions, per 1,000 students

Commission Service or Political Office/Mandate Contracts also follow the trend observed in the previous illustrations. Regions with larger populations stand out the most for this type of contract. Additionally, all regions experience a decline in numbers between 2012 and 2016, as shown in Figure 3.14.

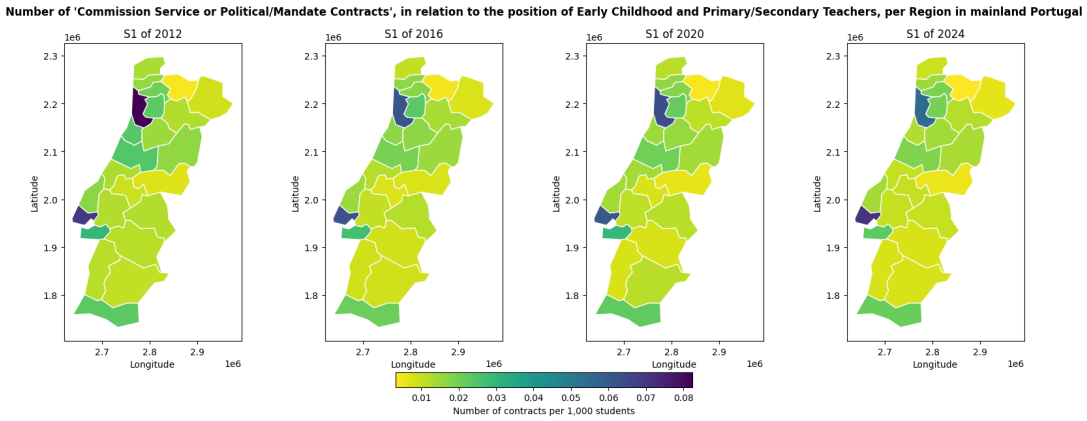


Figure 3.14. Evolution of **Commission Service or Political Office/Mandate Contracts** per NUTS III Regions, per 1,000 students

This pattern across all types of contracts may be partially explained by the population decline observed in these regions in recent years. The reduction in teacher supply between 2011 and 2015 corroborates the previously mentioned teacher shortages following the Troika period in Portugal, as discussed in Section 3.2.1.

3.2.5. Additional Features Data

In order to better understand the role of the additional features in the forecasting process, an exploratory analysis is conducted on their historical and projected values. This analysis focuses on examining the variables' behaviour over time and their relationship with the target feature, the number of positions in the public education workforce.

By comparing real and projected distributions, except for Minimum Wage and Registered Students, which only contain historical data, the analysis aims to evaluate the potential of the projected data to accurately inform future forecasts, while also assessing whether any discrepancies between historical and projected values are significant. The distributions of the variables are described in Figure 3.15.

The examination of the temporal evolution of the variables reveals several patterns when contrasting real and projected values. The projections of the annual variation of GDP predominantly replicate the overall trajectory of the historical series. However, they fail to capture the sharp decline observed in 2020, the year in which the COVID-19 pandemic emerged in Portugal. This deviation could be explained by the unexpected nature of the pandemic, which could not have been anticipated in previous forecasts. Apart from this exceptional event, the projected variables generally follow the trends observed in the historical data. In the case of Public Debt Value, the real series exhibits a downward trajectory after 2020, whereas the projections maintain higher levels for a

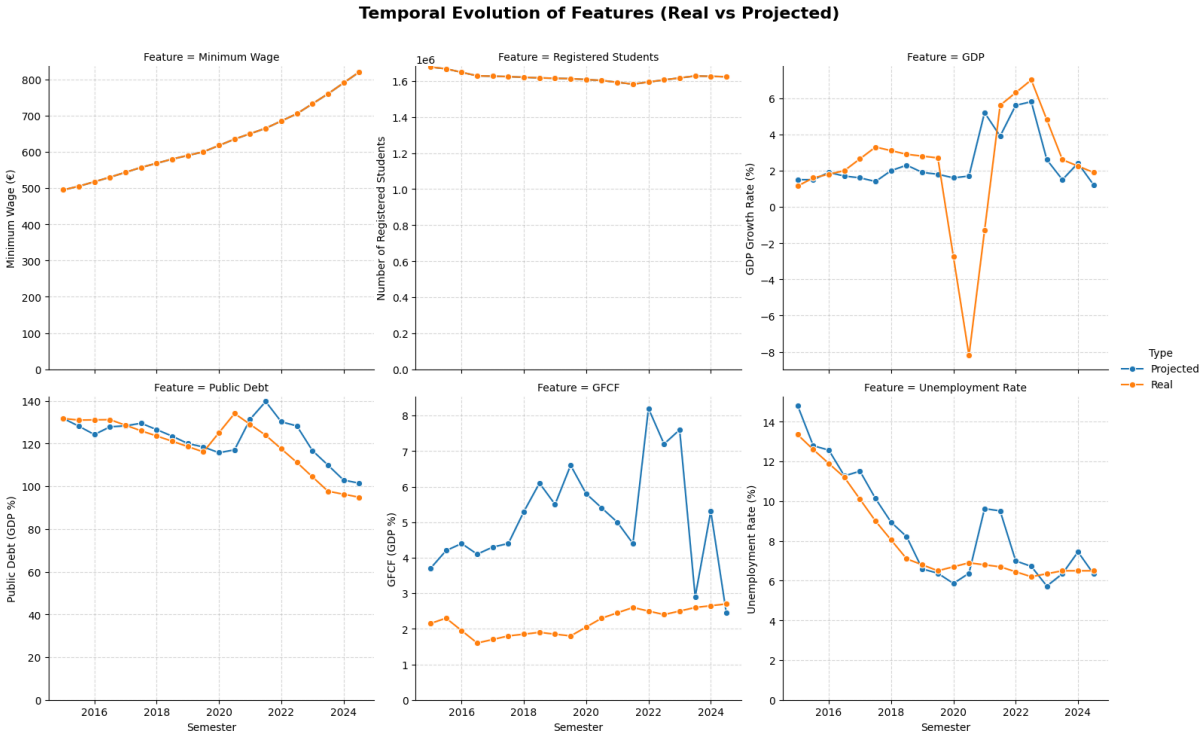


Figure 3.15. Temporal Evolution of Additional Features

longer period. This overestimation points to a tendency of the model to assume greater persistence in debt levels than is actually observed. For GFCF, also denominated Public Investment in the study, the divergence is more substantial, as real values remain relatively stable and low, with only one unexpected rise in the year 2020. In contrast, projected values are consistently higher, presenting several peaks. This suggests that the projections incorporate expectations of more aggressive investment cycles than those implemented in reality. Finally, regarding the Unemployment Rate Value, both series follow a declining trend, but the projections generally overestimate unemployment, especially in the years after 2020. This analysis highlights the discrepancies between actual and projected values, showing that all variables are affected during the COVID-19 pandemic and the subsequent years, events that earlier projections could not reasonably anticipate.

Following this comparison, the relationship between each variable and the target feature is examined through scatter plots for each dataset configuration. This step aims to provide an initial visual assessment of potential correlations, patterns or behaviours that may influence model performance in the forecasting stage, as well as support assumptions discussed later in the results.

The ministries dataset, along with the three datasets for different teaching positions, indicates that the projected additional variables, Minimum Wage, Number of Registered Students and Unemployment Rate have a stronger influence on the target feature compared to the other additional variables. Figure 3.16 illustrates the relationship between these additional features and the target variable for the dataset concerning Polytechnic

Higher Education Teachers. The other datasets, when not distributed by Age Group, follow a similar pattern of results.

When analysing the real additional features, presented in Figure 3.17, a similar trend can be observed. However, in this case, the correlation between the number of students and the target variable is inverse and less pronounced. In contrast, GFCF and Public Debt exhibit stronger and more direct correlations with the target feature than in their projected form.

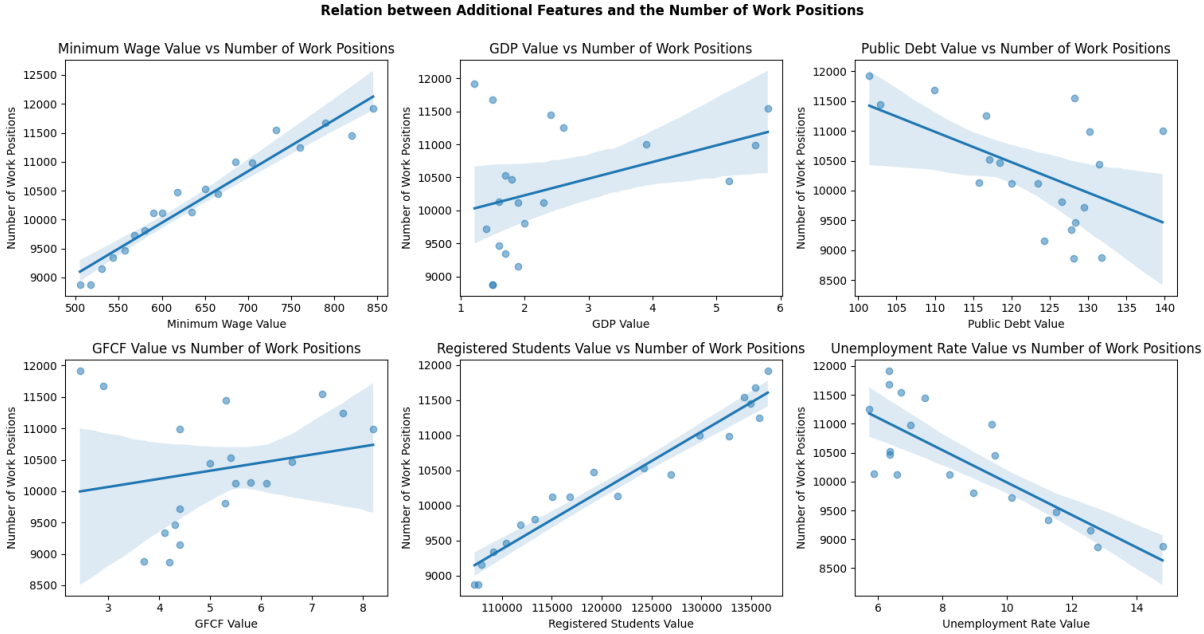


Figure 3.16. Relation between **Projected Additional Features** and the Target variable, the number of Work Positions, for **Polytechnic Higher Education Teachers**

The results indicate a strong positive relationship between these features and the number of work positions, as evidenced by the close alignment of data points with the regression line. For projected features, specifically, higher values of Minimum Wage and Registered Students correspond to a greater number of work positions. In contrast, the Unemployment Rate exhibits an inverse relationship, where lower unemployment is associated with a higher number of employees. These findings highlight these variables as the most relevant for inclusion in the forecasting phase of the study, given their closer association with the target feature.

The datasets segmented by Age Group do not exhibit a uniform pattern, as each age group shows different features with stronger correlations to the target variable. Additionally, the relationships between features and the target are less consistent, with fewer plots showing data points closely aligned to the expected regression line.

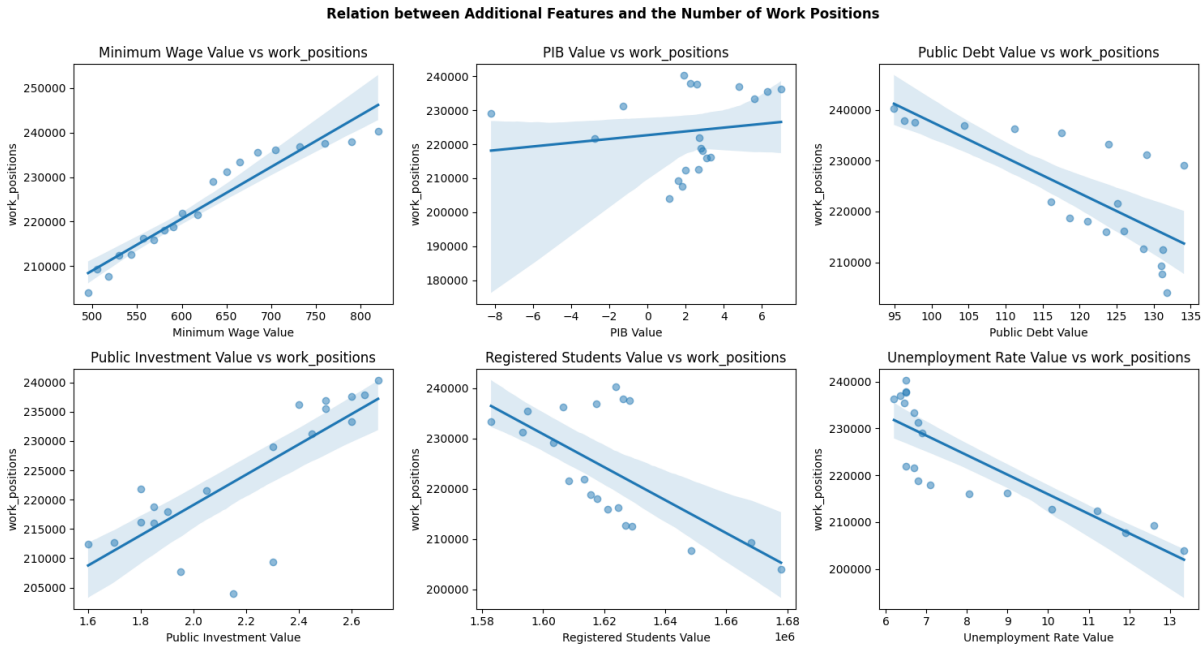


Figure 3.17. Relation between **Real Additional Features** and the Target variable, the number of Work Positions, for the **Education Ministry**

CHAPTER 4

Methodology

Firstly, it is important to note that the collected data is organised into four distinct datasets: one that aggregates information from complete ministries, and three others corresponding to specific teaching categories. Each dataset is used independently to produce separate forecasts, to predict biannual values specific to each category.

Additionally, to simplify the analysis, only data from mainland Portugal is considered, described as the Central Administration, therefore excluding the autonomous regions of Madeira and Azores. Due to the discrepancy in flow patterns before and after 2015, described in Section 3.2, the initial phase of the analysis will be conducted using two datasets: the complete dataset and the subset comprising only data from 2015 onward. The dataset exhibiting superior performance, indicated by a lower RMSE value, will be selected for use in the subsequent stages of the methodology.

The overall process is guided by the CRISP-DM methodology, serving as a general framework for the development of the analysis and forecasting approach. In the first step, the source data and additional features are extracted and structured to improve clarity and organisation and to prepare the data for the following stages. For the main datasets, only records related to the education sector are extracted, and the data is organised by semester and year. For the additional features, only the selected variables are retrieved from the source datasets. In the case of projected features, only values corresponding to the specific publications and years of interest are considered.

No data cleaning or outlier treatment is performed, as the data consists of direct observations with one value per variable for each semester. Given that the data points are independent and aggregated at the semester level, typical issues such as missing values, duplicates or outliers are not present or relevant in this context. Since the historical data for additional features is provided on an annual basis, it is converted to a semi-annual format through interpolation. Available values are used for the first semester of each year, while the median of these values is employed to estimate missing periods.

All prepared data is then combined into two distinct datasets: one combining the main dataset with projected feature values and another combining it with real historical values. These two datasets support the dual-approach described in Section 4.2.

Following Exploratory Data Analysis, the structured data is used to forecast the number of employees in the Portuguese education sector using the procedures mentioned in Section 4.1. The predictive performance of the different models is then evaluated using the selected performance metrics. Ultimately, the best-performing model, defined by

the lower RMSE metric value for each forecasting category and age group, is selected to generate predictions for future semesters.

4.1. Forecasting Approaches

Inspired by the dual approaches described in Section 2.4.3, both TS methods and ML algorithms are applied and evaluated to assess their effectiveness in forecasting public education data. Therefore, the forecasting process is structured into three distinct approaches: two baseline simulation models, a TS approach and a ML application. Each of these is described more thoroughly in the next sections.

For the predictions, the data is aggregated by age group only, without distinguishing between genders. This means that for each category and record, the values for males and females are summed, and the resulting total is then associated with the corresponding age group.

4.1.1. Baseline Simulation Models

Simulation 0 is a basic model that forecasts workforce positions based solely on the ageing phenomenon. It assumes that, each semester, a fixed share of workers progress from one age group to the next, proportionally distributed according to a delta parameter. For example, it is assumed that 5% of the workforce in a given age group transitions to the subsequent group each semester, while 95% remains within the same group. The youngest age (under 24) group does not receive any inflow, whereas the oldest group (over 65) is subject to an additional 5% reduction to account for retirements and other exits. This model establishes a baseline forecast, excluding external influences such as recruitment or voluntary departures.

Following this, a second simulation, called Simulation 1, incorporates variability by using the mean and standard deviation of the deviations observed in Simulation 0 to define a normal distribution. Random values are sampled from this distribution to simulate potential additional entries and exits from the workforce.

The outcomes of these simulations serve both as a baseline for subsequent forecasting methods and as a benchmark to compare metric values, providing insight into the relative advantages and limitations of each approach.

4.1.2. Time Series Approach

TS forecasting is a statistical technique used to estimate future values by analysing historical data and identifying underlying temporal patterns. In parallel, ML methods can be applied to learn complex relationships and hidden structures within the data, often improving prediction accuracy beyond traditional models.

In the TS context, different models are selected for the study, including Exponential Smoothing, VAR, ARIMA and SARIMAX.

Exponential Smoothing is a technique for forecasting univariate TS data that relies on the idea that predictions are computed as weighted linear combinations of past observations (lags), efficiently identifying trends and seasonality in the data.

VAR is a statistical model that captures the interdependencies among multiple TS by modelling each variable as a linear function of its own past values and the past values of all other variables in the system.

ARIMA is a widely used statistical model for TS forecasting, particularly effective in handling non-stationary data by differencing and capturing autoregressive and moving average components to model temporal dependencies. Its relative simplicity of implementation makes it attractive across multiple domains. For instance, ARIMA models have successfully provided accurate forecasts in food manufacturing, where an ARIMA(1,0,1) configuration proved effective in predicting future demand and offering reliable guidance for decision-making in this sector [33].

Despite its traditional formulation not accounting for external influences, the model's flexibility is extended through the integration of exogenous regressors, resulting in a model described as Autoregressive Integrated Moving Average with Exogenous Regressors (ARIMAX). This allows for the incorporation of external explanatory variables, thereby enhancing the model's capacity to capture complex dynamics influenced by socioeconomic or structural factors beyond the historical values of the target series. For example, [34] applies both ARIMA and ARIMAX models to predict cooking oil prices, using domestic fresh fruit bunch prices and a COVID-19 pandemic indicator as exogenous variables. The ARIMAX model achieves a MAPE of 17.31%, compared to 17.69% for the ARIMA model, demonstrating a modest but important improvement in forecasting performance.

To address seasonality, SARIMA extends ARIMA by incorporating seasonal terms. Similarly, SARIMAX extends ARIMAX to support both seasonality and exogenous variables simultaneously.

These models are selected due to their interpretability, robustness and proven effectiveness in modelling structured temporal patterns.

The incorporation of exogenous regressors enables a comparative analysis between the ARIMA and SARIMA models and their extended versions, ARIMAX and SARIMAX, respectively, without and with the inclusion of additional external features, as detailed in Section 3.1.2. This facilitates an evaluation of the extent to which these external variables enhance the models and quantifies the impact on their performance capabilities.

The initial step of the TS procedure involves evaluating the models on historical data by partitioning the dataset into training and testing subsets according to a selected training proportion. The model predictions on the testing set are then compared to the actual observed values to assess prediction error, and the training proportion yielding the best performance is selected. This evaluation process is conducted for each of the four datasets, both at an aggregated level and disaggregated by age group, thereby enabling the forecasting of workforce positions within each specified age category.

4.1.3. Machine Learning Application

For the ML approach, the selected models include Random Forest Regressor, SVR, Gradient Boosting Regressor, XGBoost Regressor [35] and MLP Regressor.

Random Forest Regressor is an ensemble method that combines predictions from multiple decision trees to enhance prediction stability and reduce variance. The SVR model aims to identify a function that best approximates the relationship between input features and a continuous target variable while minimising prediction errors.

Gradient Boosting and XGBoost build strong predictive models by sequentially combining many simple learners, each correcting the errors of its predecessors.

MLP is a type of artificial neural network consisting of multiple layers of interconnected neurons, which apply non-linear activation functions, enabling the model to capture complex and non-linear patterns within the data.

These models are capable of detecting complex, non-linear relationships in the data, providing an advanced technique for forecasting future values.

The evaluation methodology previously described for the TS models, that is, splitting the data into training and testing sets, is also applied here.

4.2. Model Enhancements and Feature Integration

To further improve forecasting performance, a set of additional features is incorporated, as described in Section 3.1.2:

- GDP (Gross Domestic Product)
- GFCF (Gross Fixed Capital Formation)
- Public Debt
- Unemployment Rate
- Minimum Wage
- Number of Registered Students

The selection of additional features is guided by their potential to reflect macroeconomic, social and demographic factors that influence the dynamics of public sector employment, particularly within the education sector.

GDP and GFCF provide information about the country's economic strength and investment levels. Public Debt and Unemployment Rate help to understand government budget constraints and labour market conditions. Minimum Wage is related to salary policies and hiring capacity, and, finally, the number of Registered Students is directly linked to staffing needs in the education sector.

The feature corresponding to the number of Registered Students is segmented according to the level of education. When forecasting positions for university teaching staff, only the number of students enrolled in universities is considered, and the same logic is applied to the other teaching categories. For forecasts at the ministry-wide level, the total number of students across all education levels is used.

Following this refinement, the models are evaluated by testing all possible combinations of these additional features, alongside different training–testing splits, selecting the configuration that produces the best predictive performance.

As these features are not available beyond the forecast horizon, the evaluation is carried out in two phases: first, using the actual historical values for the training period; second, using the projected values for the same period. This approach allows an assessment of how discrepancies between real and projected data may affect the model's accuracy and whether the models remain viable when relying on forecasts rather than observed data.

The number of Registered Students is constant across both trials, as this variable can typically be projected with relatively low error, being largely based on historical birth rates and previous enrolment trends. A similar assumption is made for the Minimum Wage, which typically follows a stable and predictable trajectory. For the remaining features, which tend to exhibit greater variability, the forecast values used are generated for the year immediately preceding each prediction. In the case of forecasts beyond 2024, projections for two or three years ahead are used, depending on the availability of data. In the case of real values of additional features, data is only available until 2024, which prevents interpolation of the second semester of that year. To address this issue, annual values are used to define the second semester, while the first semester is interpolated instead. This approach allows the use of 2014 values to interpolate the first semester of 2015, ensuring no limitations in the available data or in the assessment of the models.

In summary, the objective of this step is to analyse the impact of incorporating external explanatory variables and assess the performance of each model configuration.

4.3. Evaluation Metrics

To assess the performance of each model, several metrics are considered, including RMSE, Mean Squared Error (MSE), Mean Absolute Error (MAE) and MAPE, as all of them provide insights into model evaluation, with lower values indicating better predictive performance. However, only RMSE is employed in this study as the evaluation metric, as it penalises larger errors more significantly, thereby offering a more sensitive measure of prediction accuracy, especially for outliers and larger deviations from actual values.

To ensure a fair comparison of model performance across different age groups, an additional metric is introduced: the Normalised Root Mean Squared Error (NRMSE). This metric standardises the RMSE by dividing it by the mean of the actual observed workforce positions within each age group, and is defined as follows:

$$\text{Normalised RMSE} = \frac{\text{RMSE}}{\overline{y}} \ , \tag{4.1}$$

where $\overline{y}$ denotes the mean of the actual observed values of workforce positions. This normalisation enables a direct and fair comparison of forecasting errors across age groups of different sizes, ensuring that the evaluation remains balanced regardless of the absolute magnitude of each group.

4.4. Future Forecast Procedure

The collected data from the previous steps is then used to forecast future work positions. The number of semesters to forecast is defined after the model evaluation, with basis on the performance results, as a fixed percentage is used for all examples, to ensure consistency of forecasts.

The forecasting process follows the same modelling procedure as in the training phase to predict the future education workforce. It begins by gathering projections of additional features for the forecasting period and integrating them with the available data. Since projections for the Unemployment Rate and Public Debt in 2027 are not available from the OECD, the source used for these features during the modelling phase, values from *Banco de Portugal*'s projections [30] are used instead. Minimum Wage and the number of Registered Students, which are treated as real values during the modelling phase, are defined for the first time in this phase. Projections for the Minimum Wage are obtained from Portugal's government news portal, which provides forecasts up to 2028, available at [36]. As no official projections exist for the number of Registered Students, disaggregated by different teaching positions, these are estimated using a calculated Annual Growth Rate (AGR), defined as follows:

$$\text{AGR} = \left(\frac{V_{\text{final}}}{V_{\text{initial}}} \right)^{\frac{1}{\text{time periods}}} - 1 \ , \tag{4.2}$$

where V_{initial} is the initial number of students available (beginning of 2015), V_{final} is the final number (end of 2024), and time periods is the number of temporal intervals, in this case, semesters, between them. This approach provides a straightforward yet effective method for forecasting, as it is easy to implement while still yielding robust estimates. Studies indicate that, in domains such as hazardous waste production, it can achieve smaller deviation ranges than alternative models, underscoring its reliability when sufficient historical data is available [37]. Accordingly, the formula is applied to each forthcoming semester to estimate the corresponding number of students.

The AGR values for each category are as follows:

- Early Childhood Educators and Primary/Secondary Teachers: -0.88%,
- Polytechnic Higher Education Teachers: 1.76%,
- University Teachers: 2.58%.

For the Education Ministry category, no separate AGR exists, as its projected values correspond to the sum of the three categories above.

Following the modelling phase, the most suitable forecasting method is identified and subsequently employed to generate future projections. To ensure both consistency and rigour, the modelling process is repeated across all categories using a fixed temporal horizon, thereby aligning evaluation with the intended forecasting period. The model and feature combination that achieves the best performance, as measured by the lowest RMSE, is then selected and used for forecasting, with only the relevant exogenous

variables retained in the final predictions. Using the complete dataset alongside the projected additional features, forecasts are produced for each future semester, both for the aggregated dataset and for each Age Group subset, and the results are reported.

The same procedure used for model evaluation with a fixed number of semesters and future forecast predictions is also applied to forecast the number of contracts by Type of Contract, based on the dataset described in Section 3.1.1.

4.5. Tools and Development Environment

The analysis and predictive model implementation are carried out using *Python* version 3.11.5 [38], with Jupyter Notebooks running in the *Visual Studio Code*[1] environment, for TS approaches, and *Google Colab*[2] for ML methods, due to the increased resource requirements and substantial computational cost of the MLP model.

The main libraries used include *pandas* [39] and *numpy* [40] for data manipulation and analysis; *openpyxl* for text processing and structuring of source data [41]; *geopandas* for spatial analysis of the data [42]; *statsmodels* for statistical modelling [43]; *XGBoost* [35] and *pmdarima* [44] for specific modelling tasks; *scikit-learn* for building and validating ML models, as well as evaluating TS and ML methods [45]; *matplotlib* [46] and *seaborn* [47] for data and results visualisation; *itertools* for generating parameter combinations during the search for the best-performing model [48].

4.6. Model Implementation

In this section, the model's implementation is described, together with the description of the modelling process and the specification of the technicalities and parameters of each model used.

The dataset is organised on a semi-annual basis, with forecasts carried out separately for each age group and for each dataset type. The target variable in each case is the number of employees in the education sector for a given semester. To determine the most suitable training-test split, multiple proportions are tested, ranging from 60% to 80% of the data for training, in increments of 5%. This approach is applied consistently across both TS and ML models. In each case, the most recent semesters within the chosen split are used as the test set to evaluate model performance on future data.

4.6.1. Time Series Models

The TS models are applied using combinations of model type and training-test split. For each case, the data is split according to the chosen ratio and the model is run using reusable functions that adapt based on the model and whether exogenous variables are included.

For models with seasonality, such as Exponential Smoothing, SARIMA and SARIMAX, the seasonality is set to 2, reflecting the semi-annual nature of the data. Exponential

[1]Microsoft code editor, available at `https://code.visualstudio.com`.
[2]Hosted Jupyter Notebook service, available at: `https://colab.google`.

Smoothing, ARIMA, and SARIMA models are applied without any exogenous variables, while models like ARIMAX and SARIMAX include additional explanatory features.

The Exponential Smoothing model is configured with an *additive* trend and an *additive* seasonal component. The VAR model is fitted using the Akaike Information Criterion (AIC) to determine the optimal number of lags, which are then used to forecast based on the most recent observations. For ARIMAX and SARIMAX models, the optimal parameters are automatically selected using the ADF test to assess stationarity. The parameter selection is implemented using the *pmdarima* library, which automates the identification of the best model configuration, to ensure appropriate model fitting. The same procedure is applied to the ARIMA and SARIMA models. This approach ensures a consistent implementation of each model's specifics and data structure.

4.6.2. Machine Learning Models

For the ML models, the process differs from the TS approach. Here, the model is first defined using a pipeline, including any preprocessing steps and specific model parameters. The implementation is subsequently repeated across different training–testing split ratios to determine the optimal proportion for each model.

In this approach, the features used, including the selected additional variables and the corresponding semester, are scaled using *StandardScaler* to ensure consistent input distributions for the algorithms. No categorical features are included, so no encoding or other treatment techniques are required. In order to ensure consistency and reproducibility of the results, all models are initialised with a fixed random state, where applicable. The XGBoost Regressor is configured with 10 estimators, and the MLP Regressor is defined with two hidden layers of size (50, 50).

The MLP model is trained on the total sum of positions of the Education Ministry, for each combination of additional features and each percentage of the train–test split. For every run, the corresponding loss curve is analysed to determine the most suitable number of maximum iterations. Convergence is typically observed between 2,500 and 12,000 iterations, with the worst-case scenario requiring just under 25,000 iterations to achieve optimal performance, as described in Figure 4.1. Consequently, the maximum number of iterations is fixed at 25,000 for all subsequent model experiments.

Both approaches are supplemented with additional features to enhance model performance. These features are chosen and combined in different ways, and for each combination, the procedures described above are applied. For each category and age group, the specific selection of additional features, model and training-testing split that yields the best RMSE evaluation value is selected. These best-performing setups are then used to generate forecasts for the designated future semesters.

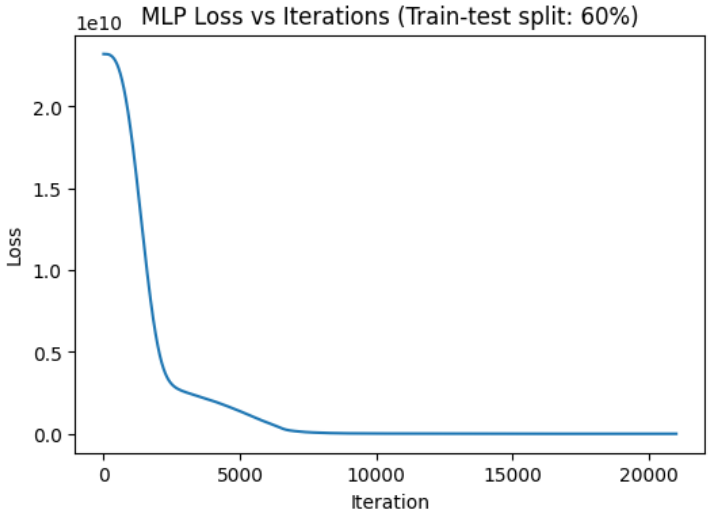


Figure 4.1. MLP Worst Loss Curve

[This page is intentionally left blank.]

CHAPTER 5

Results

This section presents the results obtained through the methodology outlined in the preceding chapters. The baseline models (Simulations 0 and 1) serve as reference points for the forecasting process, relying exclusively on historical data of work positions. In contrast, the TS and ML Approaches represent more advanced methodologies, as they integrate additional social and economic variables into the forecasting process.

The data in the plots is displayed by semester, with each full year marking June and the intermediate points marking December.

5.1. Comparison of Full Dataset vs. Post-2015 Subset Forecast Performance

Given the significant shift in workforce trends observed when comparing the dataset periods before and after 2015, available at Figure 3.3, a simple TS model is defined and evaluated under two conditions: using the entire available dataset (2011 to 2024) and using only data from 2015 onward.

This model is defined using only the work positions data from previous years, not using the specified additional features. The model uses the same structure as the predictive model, exploring the combination of percentage and model that accounts for the best predictive result. The results of both conditions are described in Figure 5.1, covering the overall Ministry of Education. The plot shows the comparison between the real data and the values from the two described forecasting experiences. These results demonstrate an improvement from the first to the second condition, with the RMSE decreasing from 2,212.46 to 1,783.11.

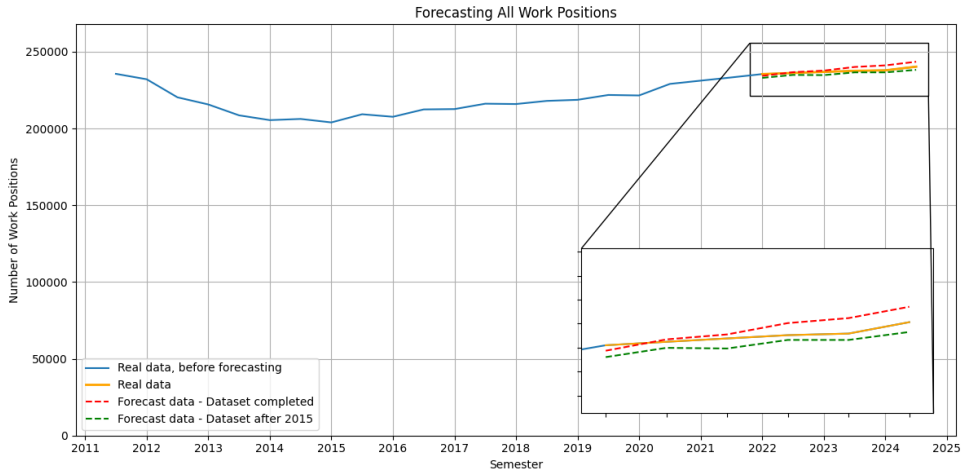
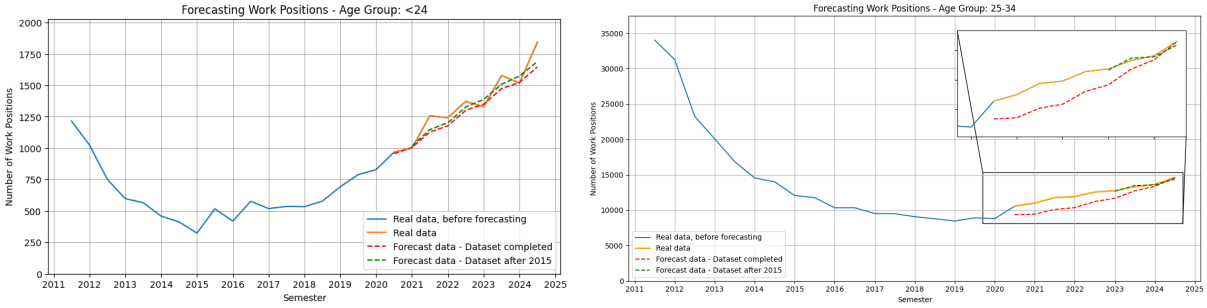


Figure 5.1. Results of Simple Model using Dataset Completed and using only Data from 2015 onward, for the **Education Ministry**

The work positions within the same ministry, specifically for the age groups under 24 and from 25 to 34 years old, are described in Figures 5.2a and 5.2b, respectively. Their RMSE values range from 92.19 to 80.05 for the first age group and from 1,180.41 to 154.01 for the second age group.



(a) Age Group: <24 **(b)** Age Group: 25-34

Figure 5.2. Results of Simple Model using Dataset Completed and using only Data from 2015 onward, for the **Education Ministry** and divided by Age Group

The difference in the starting semester of both projections is due to discrepancies in the optimal training-testing split of the best model for each approach.

These results reveal a substantial difference in RMSE values between the two scenarios, reflecting the significant economic and social changes in Portugal during that period and their consequent impact on the education sector. Since 2015, workforce trends have stabilised, exhibiting consistent patterns in the subsequent years. Accordingly, the subset of data from 2015 onward is selected for the forecasting process, as it yields better results and more accurately reflects the trends observed in workforce dynamics, compared to the full dataset.

5.2. Forecasting Approaches Results

This section presents the results of the tested forecasting approaches, providing a description of the outcomes and an interpretation of the findings. The analysis focuses on the completed datasets, as well as the subsets of data comprising each different age group.

For the purpose of interpreting the findings, the best numerical results for each category (ministry or teaching position) and age group are summarised in Table 5.1. The 'Total' categorisation within the Age Group section, while not representing a specific age group, aggregates data across all age segments and provides an overview of the overall evolution of workforce numbers in the sector.

5.2.1. Baseline Simulations Results

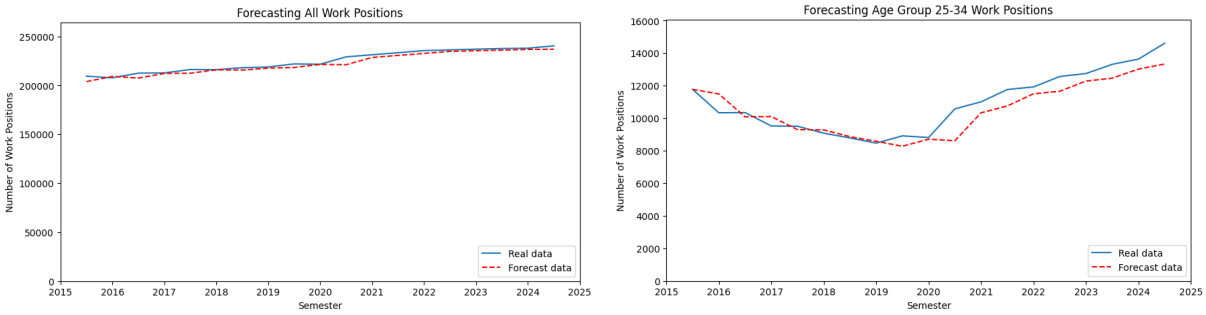
In this section, an analysis of the baseline models is presented, highlighting their characteristics and evaluating their predictive performance. These models do not incorporate additional features, relying solely on historical data to forecast future employment numbers.

Table 5.1. RMSE values (rounded to 2 decimals), per Category (ministry or position) and Age Group

Category	Model Set	<24	25–34	35–44	45–54	55–64	>65	Total
Education Ministry	Simulation 0	156.03	778.91	897.98	1,262.60	1,899.34	761.85	3,256.96
	Simulation 1	**22.53**	249.00	1,045.33	732.24	310.61	1,214.88	392.70
	Time Series Real	32.84	**115.99**	306.65	237.19	312.70	396.07	410.62
	Time Series Projected	68.51	152.46	**217.54**	**186.33**	**195.76**	**309.14**	**361.13**
	Machine Learning Real	76.62	218.99	1,255.79	1,115.39	1,156.36	483.47	2,357.49
	Machine Learning Projected	70.97	410.19	3,074.91	1,405.59	1,156.36	448.48	2,359.07
Early Childhood + Primary/Secondary Teachers	Simulation 0	77.02	339.30	1,001.85	656.71	1,274.65	510.24	1,382.50
	Simulation 1	**5.34**	68.62	645.62	517.84	233.54	756.11	247.43
	Time Series Real	42.02	44.73	249.73	**113.02**	**164.74**	243.09	319.79
	Time Series Projected	19.25	53.90	**240.54**	127.32	204.62	**226.12**	**200.77**
	Machine Learning Real	45.20	**36.57**	368.07	348.51	702.56	396.08	215.69
	Machine Learning Projected	47.69	85.92	1,181.51	346.08	702.56	242.42	292.96
Polytechnic HE Teachers	Simulation 0	17.60	78.24	138.89	122.15	58.81	30.87	322.78
	Simulation 1	**2.14**	23.76	52.03	**15.70**	37.46	37.14	**16.97**
	Time Series Real	6.90	17.42	**29.99**	22.87	23.07	13.41	52.60
	Time Series Projected	5.74	**14.55**	34.27	34.49	**19.00**	**13.07**	37.33
	Machine Learning Real	8.22	42.01	119.28	68.80	34.74	19.48	252.74
	Machine Learning Projected	3.30	44.93	119.23	111.11	115.06	21.21	214.62
University Teachers	Simulation 0	50.25	122.72	88.47	58.53	87.21	22.73	335.90
	Simulation 1	**5.98**	31.78	46.98	38.04	**11.15**	46.50	**21.00**
	Time Series Real	9.32	32.49	**7.87**	**14.63**	24.30	**13.06**	63.14
	Time Series Projected	8.79	**30.63**	27.55	16.85	17.30	15.50	44.64
	Machine Learning Real	31.20	55.15	20.53	59.11	55.32	13.17	239.41
	Machine Learning Projected	21.53	100.64	78.68	65.21	55.32	60.65	171.13

5.2.1.1. *Simulation 0*

The results obtained from Simulation 0, which provide baseline forecasts of workforce positions based solely on the ageing process, are described in Figure 5.3a for the complete Education Ministry and in Figure 5.3b for the age group between 25 and 34 years old.



(a) Age Group: Total **(b)** Age Group: 25-34

Figure 5.3. Simulation 0 Results, for the **Education Ministry**

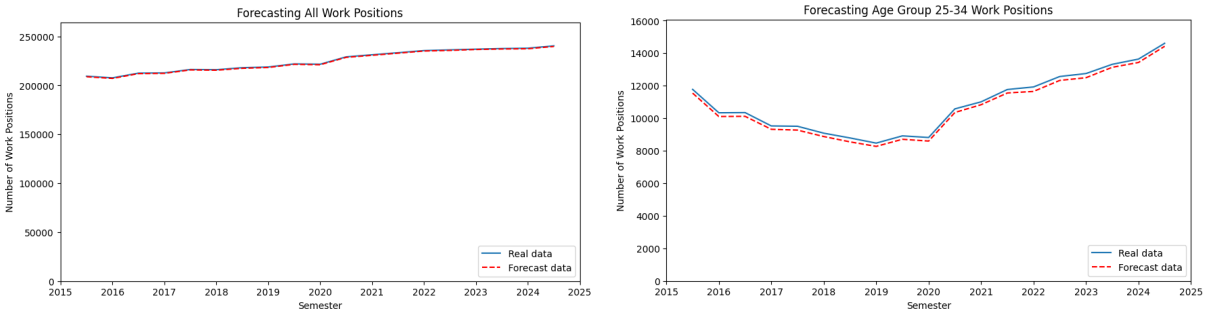
The plots compare actual and forecasted work positions between 2015 and 2024. In the first case, the model exhibits counter-intuitive behaviour in the second semester of each year, until 2020, with forecasts declining while real values increase, indicating weaker performance towards the end of each year. A clear divergence emerges in December 2020, when real data rises sharply, but the forecast model does not follow this change. From 2021 onward, the model captures the general upward trend more effectively. In general, the forecasted data follows the same trajectory as the real data, reflecting its overall growth.

In the second plot, the same tendency of forecasts decreasing while real values increase persists. However, between late 2017 and 2019, the forecast and real data align closely,

with the model successfully capturing the downward trend. The largest discrepancy appears between 2020 and 2021, when the gap between the two series widens considerably. After 2021, the model realigns with the overall growth trend, although it consistently underestimates the number of work positions, remaining below the real values.

5.2.1.2. *Simulation 1*

Simulation 1 extends Simulation 0 by incorporating both age progression and simulated additional workforce entries and exits, introducing a delta parameter to model variability. For each age group, 150 iterations are performed, each consisting of 100 simulations, with the mean RMSE calculated for every iteration. The resulting forecasts for the Education Ministry as a whole and for the 25–34 age group are presented in Figures 5.4a and 5.4b, respectively.



(a) Age Group: Total **(b)** Age Group: 25-34

Figure 5.4. Simulation 1 Results, for the **Education Ministry**

In the first plot, the forecast closely follows the real data with minimal deviations, as both series exhibit a steady, continuous growth pattern, indicating very strong model performance. The second plot, focusing on the 25–34 age group, presents a more complex scenario, as the real data displays a sharper decline followed by a strong recovery between 2018 and 2021. While the forecast captures the overall trend, it consistently underestimates the actual number of positions.

Figure 5.5 illustrates the distribution of RMSE values across training iterations for Simulation 1, which incorporates variability into workforce position forecasts. Individual RMSE values for each iteration are plotted as purple points, while the solid black line represents the mean RMSE across iterations. The mean RMSE decreases consistently across training iterations, indicating that the model progressively improves its predictive accuracy as it incorporates more data and adjusts to the variations introduced in Simulation 1, indicating increased stability. The Mean RMSE reaches a minimum value of less than 100, while the overall Mean RMSE across all iterations is 392.70.

In both these simulations, the number of work positions in each semester is calculated based on the previous semester, without any aggregation mechanism. Consequently, each forecast depends on the preceding value and is not statistically independent. As such, the results are not statistically robust, but they serve to illustrate baseline trends in workforce

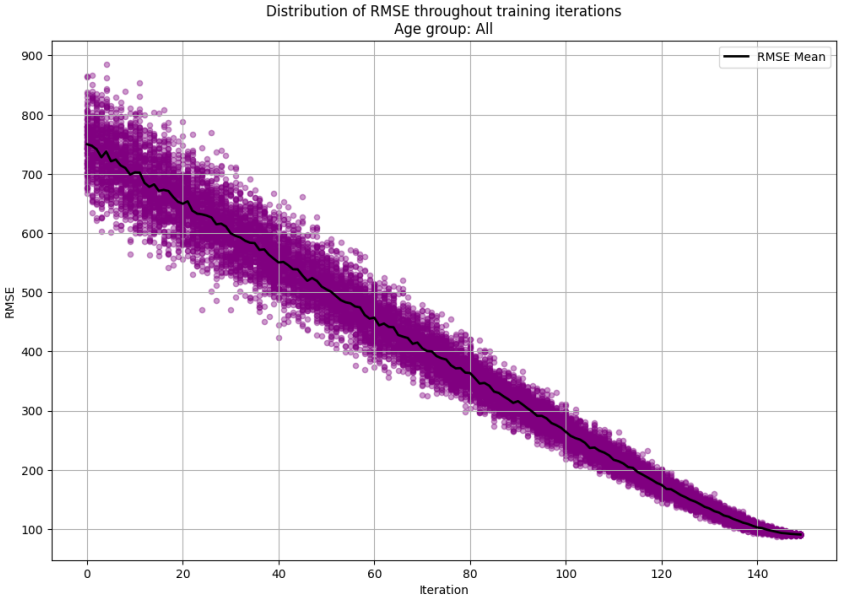


Figure 5.5. Simulation 1 Training Iterations

evolution and provide a reference for comparing more complex forecasting approaches, which results are specified in the next sections.

5.2.1.3. *Brief summary of Baseline Simulations Results*

Simulation 0 baseline consistently exhibits the highest errors across most categories, highlighting its limited predictive capability. Simulation 1 represents a substantial improvement over the initial baseline, although its performance remains uneven, with larger deviations particularly observed in the intermediate age groups.

Notably, for the prediction of total work positions in each category, Simulation 1 achieves comparatively lower RMSE values, indicating that the overall size of the education workforce remains relatively stable and does not experience large fluctuations across semesters. This stability explains why even a relatively simple forecasting approach can perform reasonably well for total counts, despite larger errors at the disaggregated age group level.

5.2.2. Time Series Approach Results

This section presents the results from the application of TS forecasting models, covering model performance across age groups and categories, evaluation of the influence of additional features and a comparison of forecasted outcomes with real observed values.

In general, TS models achieve notably lower RMSE values across nearly all age groups, highlighting their superior capacity to capture temporal patterns and trends that the baseline simulations fail to represent. These models consistently outperform the baseline simulations, particularly for the intermediate age groups where the baseline approaches exhibit larger errors. Even in cases where alternative approaches perform slightly better for specific age groups, the TS models maintain solid predictive performances overall, reflecting a well-structured and uniformly reliable forecasting strategy.

5.2.2.1. *Results between Age Groups*

Due to the varying data sizes across age groups, it is essential to evaluate their performance in relation to the corresponding sample size. To facilitate this comparison, the NRMSE metric, as defined in Section 4.3, is employed to account for these differences, and the results are displayed in Table 5.2.

Table 5.2. NRMSE values (rounded to 3 decimals) using Real and Projected Additional Features in TS Approach, per Category and Age Group

Category	**Data Type**	<24	25–34	35–44	45–54	55–64	>65	**Total**
Education Ministry	Real Data	0.022	0.009	0.007	0.003	0.004	0.030	0.002
	Projected Data	0.046	0.011	0.005	0.002	0.002	0.023	0.002
Early Childhood+Primary/Secondary	Real Data	0.097	0.012	0.013	0.002	0.003	0.033	0.002
	Projected Data	0.042	0.014	0.012	0.003	0.004	0.030	0.002
Polytechnic HE Teachers	Real Data	0.055	0.016	0.011	0.005	0.008	0.038	0.005
	Projected Data	0.045	0.013	0.012	0.008	0.007	0.037	0.003
University Teachers	Real Data	0.027	0.015	0.002	0.003	0.005	0.011	0.004
	Projected Data	0.025	0.014	0.008	0.003	0.003	0.012	0.003

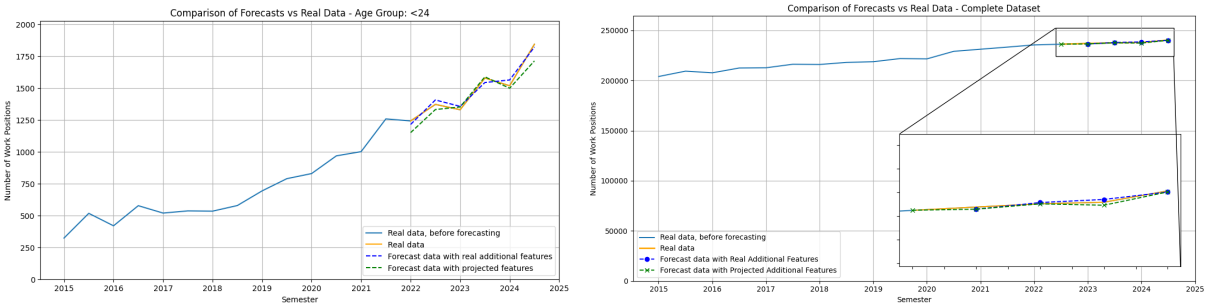
The table compares the NRMSEs obtained using real and projected additional data, for different categories of the Portuguese education workforce. Overall, the results indicate that the forecasted estimations closely match the actual values, with only minor deviations.

The youngest and oldest age groups (under 24 and above 65, respectively) exhibit the largest deviations, likely reflecting the higher variability in hiring and retirement patterns. For the youngest group of the Education Ministry category, the differences between actual and forecasted work positions, using both real and projected features, are shown in Figure 5.6a. These groups display NRMSE values in the hundredths. In contrast, the intermediate age groups show more stable employment patterns over time, typically resulting in NRMSE values in the thousandths or, occasionally, in the low hundredths.

When analysing the aggregated datasets, which represent the total number of positions across all age groups, the NRMSE values are consistent with those observed in the intermediate age groups, as illustrated in Figure 5.6b, showing the results for the Education Ministry. This indicates that the forecasting methodology maintains robustness across the full workforce distribution. In this case, the differences between using real or forecasted additional features are almost imperceptible, as their predicted values are very similar.

The worst-performing case concerns teachers under 24 years old, for whom the RMSE using real data is close to 0.1, as shown in Figure 5.7, highlighting the challenges in accurately forecasting work positions for this age group.

These observations indicate that the forecasting methodology captures overall workforce trends effectively, with only minor variations. Across all categories, the NRMSE values remain minimal, demonstrating that the method provides a highly reliable and



(a) Age Group: <24 **(b)** Age Group: Total

Figure 5.6. TS Results, for the **Education Ministry**

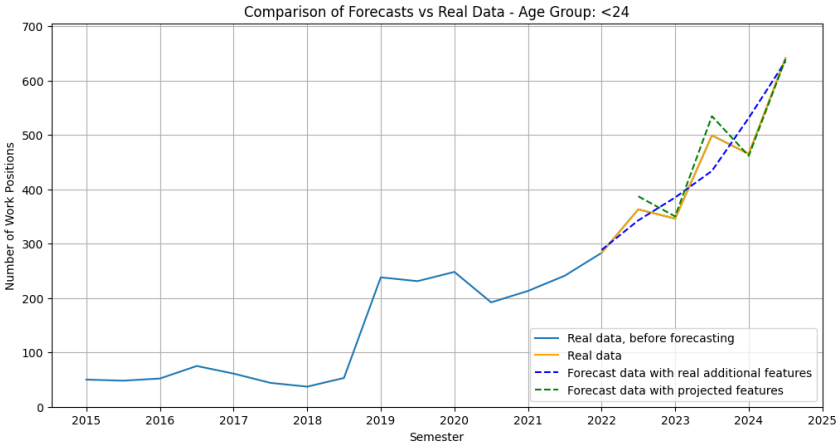


Figure 5.7. TS Results, for **Early Childhood Educators and Primary/Secondary Teachers** and divided by Age Group: <24

robust approach for predicting the number of work positions in the Portuguese education sector, both when using real and projected values of additional features.

As discussed in Section 5.1, the difference in the starting semester between the two projections arises from discrepancies in the optimal training-testing split of the best model for each approach.

5.2.2.2. *Importance of Additional Features*

In order to better evaluate the effect of incorporating exogenous features, the RMSE values obtained with the ARIMA and SARIMA models, as well as their corresponding extensions with exogenous features, ARIMAX and SARIMAX respectively, are reported for each category and age group. Table 5.3 presents the results for the ARIMA model and its extended versions using both real and additional feature values, alongside the corresponding SARIMA outcomes.

The results underscore the importance of incorporating additional exogenous features, as ARIMA and SARIMA consistently fail to outperform their extended versions in terms of predictive capability. Furthermore, the ARIMAX and SARIMAX models generally demonstrate substantially superior performance compared to their counterparts that rely solely on historical values of the target variable, delivering a consistent and significant

Table 5.3. RMSE values (rounded to 2 decimals) in TS Approach, using Real and Projected Features, per Category, Model and Age Group

Category	Model	<24	25–34	35–44	45–54	55–64	>65	Total
Education Ministry	ARIMA	80.05	276.85	1,711.78	681.72	994.96	516.92	1,783.12
	ARIMAX (Real)	55.35	255.63	346.01	280.53	389.96	479.23	410.62
	ARIMAX (Projected)	73.56	154.90	**217.54**	**186.33**	325.95	**377.44**	552.24
	SARIMA	85.51	259.82	1,587.15	2,120.62	706.43	560.00	3,519.40
	SARIMAX (Real)	**32.84**	**115.99**	343.66	697.91	542.80	396.07	1,448.59
	SARIMAX (Projected)	68.52	152.46	358.06	289.14	**195.76**	458.23	**361.13**
Early Childhood + Primary/Secondary	ARIMA	63.14	174.63	918.21	492.08	599.31	400.99	1,661.68
	ARIMAX (Real)	52.08	**48.72**	249.73	**175.47**	222.60	361.07	353.07
	ARIMAX (Projected)	**19.25**	53.90	**240.54**	184.77	271.25	234.33	649.28
	SARIMA	48.11	559.25	1,555.19	800.24	1,125.10	400.99	1,322.23
	SARIMAX (Real)	42.02	66.24	566.84	277.08	247.17	267.97	726.19
	SARIMAX (Projected)	52.27	179.58	356.83	192.68	**204.62**	**226.12**	**200.77**
Polytechnic HE Teachers	ARIMA	12.61	48.18	67.92	67.29	27.16	30.12	367.57
	ARIMAX (Real)	6.90	18.20	36.01	40.29	23.07	15.93	90.67
	ARIMAX (Projected)	**5.74**	**14.55**	36.12	39.73	**19.00**	17.84	**37.33**
	SARIMA	17.51	72.84	152.79	104.98	46.78	26.97	228.02
	SARIMAX (Real)	9.73	17.42	36.12	59.80	23.07	**13.41**	78.42
	SARIMAX (Projected)	7.68	18.46	**34.27**	**34.49**	19.00	16.40	60.47
University Teachers	ARIMA	29.80	65.95	178.79	57.44	169.01	20.67	266.59
	ARIMAX (Real)	9.32	34.83	**27.15**	25.50	26.61	**13.06**	97.59
	ARIMAX (Projected)	**8.79**	**30.63**	49.35	22.13	**17.30**	21.94	**44.64**
	SARIMA	13.30	42.40	151.25	19.55	64.20	27.00	223.82
	SARIMAX (Real)	11.55	32.49	46.35	**14.84**	64.21	16.70	63.14
	SARIMAX (Projected)	15.02	34.32	49.08	16.85	33.98	19.85	46.60

improvement in forecasting future employment positions. Occasionally, ARIMA and SARIMA achieve the same RMSE values, as do ARIMAX and SARIMAX, indicating that seasonal patterns are not detected in these instances.

5.2.2.3. *Difference between Real vs. Projected Additional Features*

One of the objectives of this project is not only to assess whether the proposed models can accurately forecast future workforce trends in the Portuguese public sector, but also to evaluate their effectiveness when relying on anticipated values of the additional input features. To this end, two experimental settings are considered: one using the actual values of the additional features and another using projected values that would have been available several years before the target period.

A comparative analysis of model performance reveals that, for the TS approach, the use of projected features tends to produce slightly lower RMSE values when forecasting the Education Ministry data, although the differences are not generally significant. These results are available in Table 5.2. When analysing the Total dataset, which aggregates all age groups, the results align with those of the Education Ministry, showing that projected data features yield better performance across most categories.

A notable exception is the Polytechnic Higher Education Teachers in the 45–54 age group, for which the models perform slightly better when using real data. This case represents an outlier, deviating from the generally consistent pattern observed across the other categories, as illustrated in Figure 5.8.

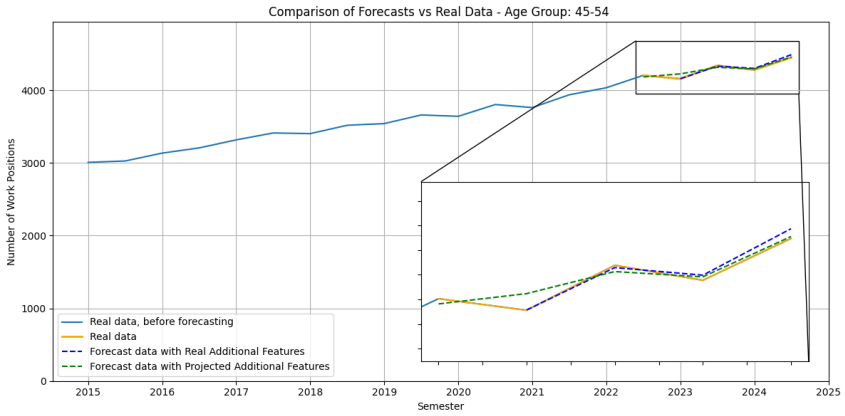


Figure 5.8. TS Results, for **Polytechnic Higher Education Teachers** and divided by Age Group: 45-54

In some cases, real data leads to better results, while in others, projected features yield lower errors. This variability limits the strength of any general conclusion, as the results indicate that the differences in model performance between the two settings are generally insignificant.

This suggests that using projected data for the additional features does not considerably compromise forecasting accuracy. Therefore, in practical applications where real future values are not available, relying on such projected inputs appears to be a viable and robust approach for anticipating workforce supply in the public sector.

5.2.2.4. *Best Model for each Category*

The best model for each category is defined in Table A.1 of Appendix A. In terms of model performance, ARIMAX and SARIMAX emerge as the most frequently effective models for forecasting the completed dataset, both achieving the best results in 3 of the 8 scenarios evaluated. It is followed by VAR, which reaches top performance in two scenarios.

When forecasts are conducted at the level of individual age groups, no single model demonstrates a clearly dominant performance. In this context, VAR emerges as the most frequently best-performing model, achieving the top position in 17 cases, followed closely by ARIMAX in 16 cases and SARIMAX in 15.

These findings indicate that the Exponential Smoothing model performs poorly in this context, while ARIMAX, SARIMAX and VAR consistently perform well, frequently competing for the best results across scenarios and age groups.

5.2.2.5. *Best combination of Features for each Category*

The number of times each additional feature is used in the best combination to forecast an age group subset or the complete dataset is shown in Table 5.4. This value is reported for each Age Group, as well as for the overall forecast, followed by the sum of all occurrences across the TS models.

Within the set of additional features employed, the Number of Registered Students stands out as the most influential variable in predicting future teacher supply, consistently

appearing in the best-performing models across all age groups. It plays a crucial role in forecasting the future workforce, being the most frequently selected external feature, with between 3 and 5 occurrences in every age group. The Unemployment Rate is the least used feature overall, with 23 occurrences. Although still relatively high, this is the lowest total among all features. Interestingly, for the age group above 65 years old, the Unemployment Rate becomes the most used feature (6 times, out of a maximum of 8). A special note is due to the Minimum Wage, which in the age group of under 24 is selected 7 out of 8 times, highlighting its importance in forecasting work positions in this subset. In contrast, in the same subset, the Unemployment Rate is used only once. The same happens for the 25 to 34 age group, where GDP appears also once, underlining their limited predictive relevance in these specific age groups.

When forecasting the aggregated workforce, the most frequently used features are the Number of Registered Students and Public Debt (4 times each), followed by the remaining features, all of which appear 3 times. This suggests a relatively balanced contribution of most features in aggregated forecasts, although the importance of students and Public Debt remains slightly higher.

For Polytechnic Higher Education Teachers specifically, the results confirm the relationships between projected additional features and the number of work positions, as discussed in Section 3.2.5. In this case, Minimum Wage and the Number of Registered Students are the features most strongly associated with the target variable. This finding is consistent with the observation that these two features constitute the optimal feature set for models at this education level, together with GDP, when using projected additional features. Regarding real additional features, the correlation patterns are also in line with previous analyses for the entire Education Ministry, as Minimum Wage, Public Debt and Unemployment Rate form the best model combinations for this category, reinforcing the conclusions drawn in the analysis of additional variables.

Table 5.4. Additional Features relevance in TS Approach, per Age Group

Feature	<24	25–34	35–44	45–54	55–64	>65	**Total**	**Sum**
Minimum Wage	4	7	4	3	3	2	3	26
GDP	6	1	5	2	2	6	3	25
Public Debt	5	2	2	3	5	3	4	24
GFCF	4	3	6	3	3	4	3	26
Number of Registered Students	5	3	5	5	4	3	4	**29**
Unemployment Rate	1	3	5	3	2	6	3	23

5.2.2.6. *Comparison of Predicted Work Positions between Total Forecast and Sum of Individual Age Groups Forecast*

In order to evaluate the consistency between the forecasts generated directly for the total number of work positions and those obtained by summing the individual forecasts of each age group, a comparative analysis is carried out. The objective is to evaluate whether

disaggregating the predictions by age group and subsequently aggregating them would yield values that align with the overall forecast obtained from the total dataset.

The comparison focused on the forecasted semesters that are common across all models and categories. For each of these periods, the difference between the aggregated age group forecast and the direct total forecast is calculated. From these differences, two summary metrics are derived. First, the MAE quantifies the average magnitude of deviation in absolute terms between the sum of the age group forecasts and the total forecast. Second, the Absolute Percentage Error measures, for each half-year, how much the aggregated age group forecast deviates from the total forecast as a percentage of the total.

Finally, the Mean Percentage Error (MPE), a metric based on MAPE, is computed as the average of these percentage errors over all forecasted semesters, indicating the average relative deviation between the two forecasting approaches. These formulas are defined as follows:

$$\text{Absolute Error}_i = |\hat{y}_{\text{sum},i} - \hat{y}_{\text{total},i}| \ , \tag{5.1}$$

$$\text{Percentage Error}_i = \frac{|\hat{y}_{\text{sum},i} - \hat{y}_{\text{total},i}|}{\hat{y}_{\text{total},i}} \times 100 \ , \tag{5.2}$$

$$\text{Mean Percentage Error} = \frac{1}{N} \sum_{i=1}^{N} \text{Percentage Error}_i \ , \tag{5.3}$$

where $\hat{y}_{\text{sum},i}$ denotes the aggregated forecast from all age groups in semester i; $\hat{y}_{\text{total},i}$ defines the direct total forecast (not age-disaggregated) in semester i; N is the number of common forecasted semesters included in the evaluation.

This analysis provides a clear understanding of how much the disaggregated forecasts deviate from the direct total forecast and helps validate the consistency and reliability of the age group approach. The results indicate that the number of forecasted positions in the education sector, when summing all age groups, aligns with the total forecast obtained without age group disaggregation, as the percentage error remains consistently low, never exceeding 1%. The MAE and MPE are calculated for each model and displayed in Table 5.5. The errors are overall very small, reinforcing the coherence of the disaggregated and aggregated forecasting approaches. The highest percentage errors are observed in the Polytechnic Higher Education Teachers category with real data, though they remain under 0.5%.

5.2.2.7. *Comparison of Total and Summed Total Forecasts with Real Values*

A comparison between forecasted values obtained from the direct projection of total work positions and those derived from the aggregation of age group forecasts, against the actual observed values in the respective semesters, enables the evaluation of which procedure provides a more accurate representation of the target variable.

Table 5.5. MAE (rounded to 2 decimals) and MPE (rounded to 3 decimals) between Summed Age Group forecasts and Total forecasts, in TS Approach

Model / Features	**MAE**	**MPE (%)**
Education Ministry (Real)	303.34	0.127
Education Ministry (Projected)	322.28	0.135
Early Childhood + Primary/Secondary Teachers (Real)	362.51	0.280
Early Childhood + Primary/Secondary Teachers (Projected)	263.86	0.204
Polytechnic HE Teachers (Real)	50.98	0.447
Polytechnic HE Teachers (Projected)	18.70	0.160
University Teachers (Real)	58.87	0.338
University Teachers (Projected)	27.75	0.158

Following the methodology described in Section 5.2.2.6, this evaluation relies on the difference between the forecasted and observed values for each semester. From these differences, the following summary metrics are derived:

$$\text{Absolute Error}_i = |y_{\text{real},i} - \hat{y}_{\text{forecasted},i}| \ , \tag{5.4}$$

$$\text{Percentage Error}_i = \frac{|y_{\text{real},i} - \hat{y}_{\text{forecasted},i}|}{y_{\text{real},i}} \times 100 \ , \tag{5.5}$$

$$\text{Mean Percentage Error} = \frac{1}{N} \sum_{i=1}^{N} \text{Percentage Error}_i \ , \tag{5.6}$$

where $y_{\text{real},i}$ denotes the real observed number of work positions in semester i; $\hat{y}_{\text{forecasted},i}$ represents the predicted number of work positions in semester i, obtained either from total forecasts or from the aggregated age group forecasts; N is the number of forecasted semesters included in the evaluation.

The results in Table 5.6 indicate that both the Total and Summed Total forecasting approaches achieve relatively low errors across all categories, although their comparative performance varies. For Early Childhood Educators and Primary/Secondary Teachers, the Total approach yields lower MAE and MPE values, suggesting that direct forecasting of aggregated series is more effective in this context. In contrast, for the Education Ministry, the Summed Total approach provides slightly better results, particularly when real features are incorporated. A similar pattern is observed for University Teachers, although in this case, forecasts based on projected features outperform those using real ones.

Overall, in the Summed Total approach, real exogenous features tend to produce slightly better results, whereas in the Total approach, projected features yield superior performance. This suggests that, when forecasting age group subsets, real values of features generate predictions that more closely align with the observed values than those

Table 5.6. MAE (rounded to 2 decimals) and MPE (rounded to 3 decimals) between Total and Summed Total Forecasts and Real Values, in TS Approach

Model / Features	**MAE**	**MPE (%)**
Summed Total		
Education Ministry (Real)	364.88	0.153
Education Ministry (Projected)	333.15	0.139
Early Childhood + Primary/Secondary Teachers (Real)	237.17	0.183
Early Childhood + Primary/Secondary Teachers (Projected)	240.08	0.185
Polytechnic HE Teachers (Real)	24.50	0.211
Polytechnic HE Teachers (Projected)	39.91	0.345
University Teachers (Real)	21.41	0.123
University Teachers (Projected)	22.22	0.126
Total		
Education Ministry (Real)	383.78	0.161
Education Ministry (Projected)	329.08	0.138
Early Childhood + Primary/Secondary Teachers (Real)	175.93	0.136
Early Childhood + Primary/Secondary Teachers (Projected)	152.75	0.118
Polytechnic HE Teachers (Real)	46.79	0.408
Polytechnic HE Teachers (Projected)	32.18	0.279
University Teachers (Real)	45.58	0.260
University Teachers (Projected)	37.72	0.216

obtained with projected features. Nevertheless, the differences are minor and, in both approaches, the errors remain modest, with percentage errors never greater than 0.7%, highlighting the general consistency and reliability of the forecasts.

5.2.3. Machine Learning Application Results

This section presents the results obtained from the application of ML forecasting models, providing an overview of their performance and highlighting key insights from their evaluation.

The ML models exhibit inconsistent performance across categories and age groups. While they occasionally outperform TS models in specific subsets, their predictive accuracy is generally less uniform and more sensitive to particular age segments. In contrast, TS models maintain more consistent performance across all age groups, highlighting their robustness, whereas ML models can leverage available features to achieve superior predictions in select cases.

5.2.3.1. *Results between Age Groups*

As previously discussed in Section 5.2.2.1, the NRMSE metric, defined in Section 4.3, is used to quantify the performance evaluation metrics differences, with results of the ML approach reported in Table 5.7.

Table 5.7. NRMSE values (rounded to 3 decimals) using Real and Projected Additional Features in ML Approach, per Category and Age Group

Category	**Data Type**	<24	25–34	35–44	45–54	55–64	>65	**Total**
Education Ministry	Real Data	0.053	0.016	0.029	0.013	0.014	0.042	0.010
	Projected Data	0.046	0.031	0.073	0.017	0.014	0.035	0.010
Early Childhood+Primary/Secondary	Real Data	0.111	0.009	0.019	0.007	0.014	0.059	0.002
	Projected Data	0.125	0.022	0.060	0.007	0.014	0.034	0.002
Polytechnic HE Teachers	Real Data	0.066	0.039	0.042	0.016	0.012	0.058	0.022
	Projected Data	0.027	0.041	0.042	0.026	0.042	0.060	0.019
University Teachers	Real Data	0.090	0.028	0.006	0.012	0.011	0.010	0.014
	Projected Data	0.062	0.047	0.022	0.013	0.011	0.048	0.010

The results reveal a clear distinction between forecasts of aggregated work positions and predictions disaggregated by age groups. NRMSE values for the aggregated dataset are consistently lower, typically in the thousandths or low hundredths, whereas forecasts by age group subsets generally exhibit higher values, often in the hundredths or even in the tenths. This pattern is particularly pronounced for Early Childhood Educators and Primary/Secondary Teachers under 24 years old, where predictions are notably less accurate, reflecting higher variability and instability, as shown in Figure 5.9. A similar pattern is observed in the TS approach, where this category and age group combination also exhibits the highest NRMSE values.

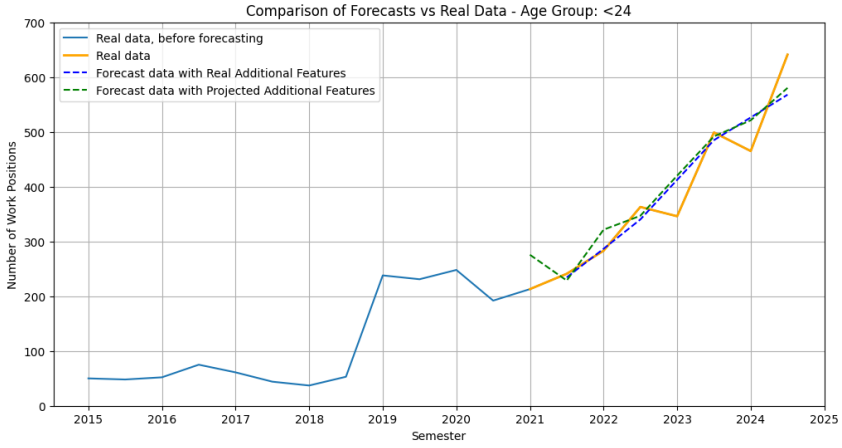


Figure 5.9. ML Results, for **Early Childhood Educators and Primary/Secondary Teachers** and divided by Age Group: <24

Overall, although the ML forecasts produce reasonable estimates, their predictive efficiency is generally inferior to that of the corresponding TS models.

5.2.3.2. *Difference between Real vs. Projected Additional Features*

Across all categories, the use of projected additional features generally results in higher RMSE values in ML models when forecasting age group subsets, highlighting the challenges of predicting workforce positions under uncertain future inputs. In contrast, for aggregated categories, which encompass all age groups, projected features often improve

performance, underscoring the distinct behaviour of the total datasets compared to the general trend observed in individual age group subsets.

For the Education Ministry as a whole and the subset of workers aged 35–44 within Early Childhood Educators and Primary/Secondary Teachers, the differences between forecasts based on real and projected features are illustrated in Figures 5.10 and 5.11, respectively.

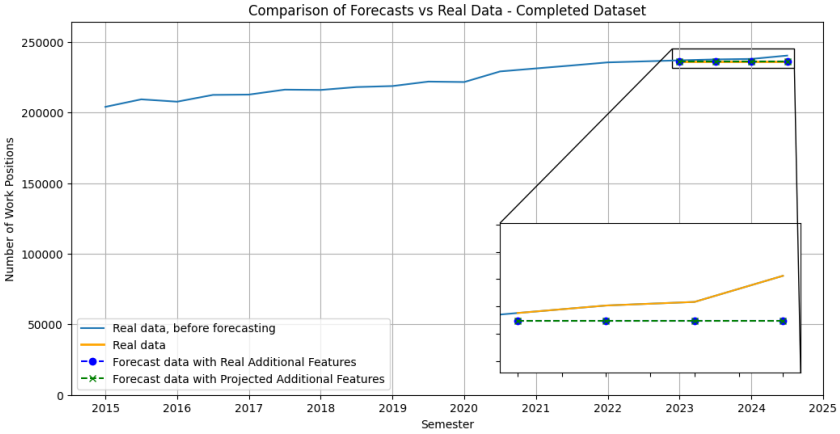


Figure 5.10. ML Results, for the **Education Ministry**

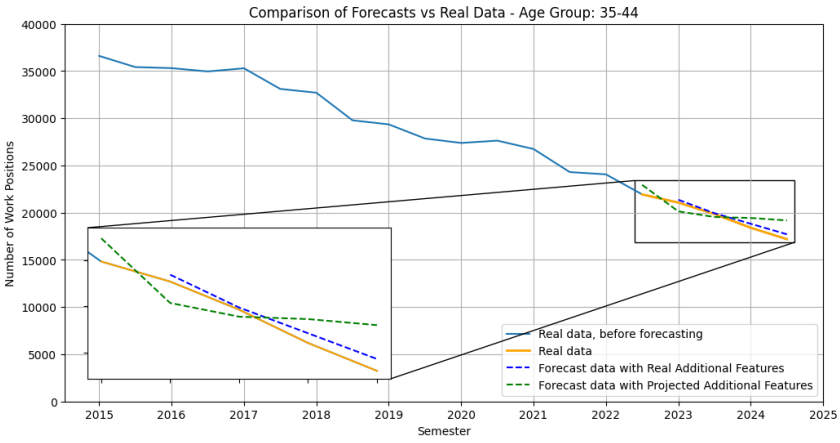


Figure 5.11. ML Results, for **Early Childhood Educators and Primary/Secondary Teachers** and divided by Age Group: 35-44

In the first case, both forecasts using real and projected additional features suggest identical results, producing a flat, linear trend, with no increases or decreases in workforce positions across the forecast horizon. In the second case, forecasts based on real features generally follow the observed trajectory, capturing some of the fluctuations, whereas forecasts using projected features produce a uniform downward trend followed by a steady linear pattern, failing to reflect the overall dynamics of the actual data. These results illustrate the discrepancies between real and projected features, corresponding to a difference in NRMSE of 0.041.

For Early Childhood Educators and Primary/Secondary Teachers aged 25–34, Figure 5.12 further highlights the closeness of forecasts to actual values, particularly when using real additional features, which align well with the observed trends.

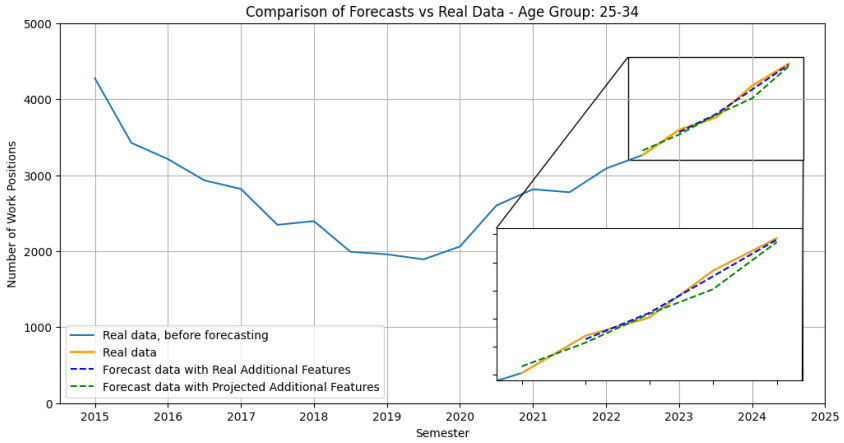


Figure 5.12. ML Results, for **Early Childhood Educators and Primary/Secondary Teachers** and divided by Age Group: 25-34

In some instances, projected features yield slightly improved results. However, forecasts based on real data generally produce lower errors. This behaviour contrasts with the TS analysis, where the contrast of performance between different feature types is less clear. Nonetheless, the differences in predictive accuracy between the two projection procedures are not substantial, making the use of projected inputs a feasible approach for forecasting workforce supply.

The linear patterns produced by the ML approach, as shown in Figure 5.10, suggest that, although the model occasionally achieves lower errors, it may not adequately capture the dynamic fluctuations of future values in this context.

5.2.3.3. *Best Model for each Category*

The best model for each category is summarised in Table A.2 of Appendix A. When forecasting the complete dataset, Gradient Boosting achieves the best results in 4 out of 8 scenarios, followed by MLP in 3 scenarios and Random Forest Regressor in 1.

For forecasts conducted by individual age groups, MLP consistently outperforms other models, being the best choice in 36 out of 48 cases. Gradient Boosting is the top model in 11 cases, while SVR only achieves the best performance for projecting University Teachers aged between 45 and 54 years old, when using projected features.

The dominance of MLP indicates that the workforce data contains complex, non-linear relationships, which neural networks capture more effectively than other models, enabling them to identify patterns across both total and age group datasets.

5.2.3.4. *Best combination of Features for each Category*

The number of times each additional feature is used in the best combination to forecast workforce positions highlights the relative importance of these features. Within the set of

additional features considered under the ML approach, the Unemployment Rate emerges as the most influential variable, being selected 32 times across all age groups and totals, as shown in Table 5.8. Its consistent presence highlights its central role in shaping workforce forecasts when using ML methods. By contrast, the Number of Registered Students is the least frequently employed feature, appearing only 14 times overall, which suggests a comparatively minor contribution to ML based predictions.

Table 5.8. Additional Features relevance in ML Approach, per Age Group

Feature	<24	25–34	35–44	45–54	55–64	>65	**Total**	**Sum**
Minimum Wage	4	5	2	1	7	2	3	24
GDP	5	2	2	1	0	4	1	15
Public Debt	3	5	2	5	1	4	3	23
GFCF	6	3	2	6	2	6	3	28
Number of Registered Students	1	3	5	2	0	1	2	14
Unemployment Rate	4	7	6	5	2	4	4	**32**

The results exhibit a clear discrepancy when compared to those observed in the TS approach, analysed in Section 5.2.2.5, where the Unemployment Rate is the least frequently used feature and the Number of Registered Students is the most frequently used. Moreover, the consistency of feature usage in the ML approach varies more markedly between features, with the lowest value being considerably smaller than in the TS models and the highest value considerably larger.

Feature relevance also varies across age groups. For instance, in the age group of between 55 to 64, Minimum Wage is selected 7 times, as well as in the 25 to 34 age group, where the Unemployment Rate dominates with 7 occurrences, reflecting their relative importance in these subsets. Conversely, certain features exhibit very limited influence, being used only once in some age groups, along with GDP when forecasting the complete dataset. In the 55–64 age group, GDP and the Number of Registered Students are never selected, indicating their unimportant relevance for predicting this subset. This pattern indicates that while ML models assign substantial weight to key predictors, their feature selection is more divergent across age groups, highlighting the complex and context-dependent nature of forecasts using ML models.

For Polytechnic Higher Education Teachers specifically, the results contrast with those of the TS models. While in the TS approach, Minimum Wage and the Number of Registered Students are the most relevant predictors, in the ML models, these two features are complemented by the Unemployment Rate, which appears with notable frequency and displays an inverse correlation with the target variable. This distinction indicates that, under the ML framework, the optimal feature set for this education level includes three variables instead of two, with the Unemployment Rate playing a more decisive role.

When considering real additional features, the divergence is again evident. Whereas the TS models highlight Minimum Wage, Public Debt and Unemployment Rate as the most relevant for the Education Ministry, the ML results point to a diminished role of

GDP and the Number of Registered Students, as these variables appear less frequently in the best-performing combinations. This pattern reinforces the conclusions drawn in Section 3.2.5, while also underlining the broader differences between the two modeling approaches.

For Polytechnic Higher Education Teachers specifically, the results confirm the relationships between projected additional features and the number of work positions, as depicted in Section 3.2.5. In this case, Minimum Wage and the Number of Registered Students are the features most strongly associated with the target variable, complemented by the Unemployment Rate, which shows an inverse correlation. This finding is consistent with the observation that these three features constitute the optimal feature set for models at this education level when using projected additional features, also reflected in TS models. Regarding real additional features, the correlation patterns are also in line with previous analyses, for the entire Education Ministry, as GDP and the Number of Registered Students appear less frequently in the best-performing model combinations of this category, reinforcing the conclusions drawn in the additional variables analysis.

In summary, labour market indicators such as Unemployment Rate, GFCF, Minimum Wage and Public Debt emerge as strong predictors of workforce dynamics, directly influencing hiring and job attractiveness. By contrast, the number of registered students and GDP show limited relevance when predicting the number of work positions in the Portuguese education sector, indicating that it does not influence teacher workforce levels in this context.

5.2.4. Summary of Forecasting Results

Overall, TS models consistently outperform the baseline simulations across nearly all categories and age group subsets, demonstrating robust and reliable forecasting performance. Notably, for the under-24 age group, Simulation 1 consistently produces the best performance values, reflecting the minimal fluctuations between semesters, within this group.

Although Simulation 1 yields favourable results, it is important to note that these outcomes are generated by introducing variation around the true values and represent an average of the best-case scenario across several simulations. As such, the results are not statistically robust, since they are not independent forecasts but rather perturbations of observed data.

ML models, in contrast, show potential in capturing complex, non-linear relationships in some subsets of the data, but their performance is less consistent across age groups and categories. Although ML approaches can outperform the baselines in specific cases, their higher variability and sensitivity limit their reliability compared with TS methods.

In summary, advanced forecasting approaches consistently surpass baselines, as TS models provide the most stable and accurate predictions for both total and specific age

group workforce positions. Consequently, TS methods are the preferred approach for projecting future work positions in the Portuguese education sector, given their consistency, reliability and superior performance.

5.3. Future Forecast

To translate the insights and results obtained from the models discussed in Section 5.2 into practical application, forecasts are generated for future semesters to estimate the number of work positions, relying on the data collected and analysed in the earlier phases of this study.

A fixed forecast horizon of six semesters is defined for all categories. This decision stems from the training experiments, where most models achieve their best performance under larger train–test splits, such as 80% or 75%. In practice, an 80% split corresponds to a testing period of approximately five semesters. However, this period does not align with complete calendar years. Extending the horizon to six semesters ensures alignment with full-year intervals, covering the period from early 2025 to the end of 2027.

Given that the TS approach demonstrates the highest effectiveness in forecasting workforce positions, it is selected as the basis for the future projections. To maintain consistency with the training–testing procedure, the modelling process is repeated using a fixed horizon of five semesters. This adjustment ensures comparability, as five semesters out of twenty correspond proportionally to six semesters out of twenty-six, thereby aligning the evaluation setup with the intended forecasting period.

For each category, the model and feature combination that achieves the lowest RMSE when forecasting the five-semester horizon during training is selected. These models are then applied to generate predictions for the test subset, incorporating only the corresponding projected values of the exogenous variables in the final forecasts. Using the complete dataset along with projected values for the additional features, forecasts are produced for each of the six semesters, both for the aggregated workforce and for the disaggregated age group subsets.

5.3.1. Future Forecast Model Performance

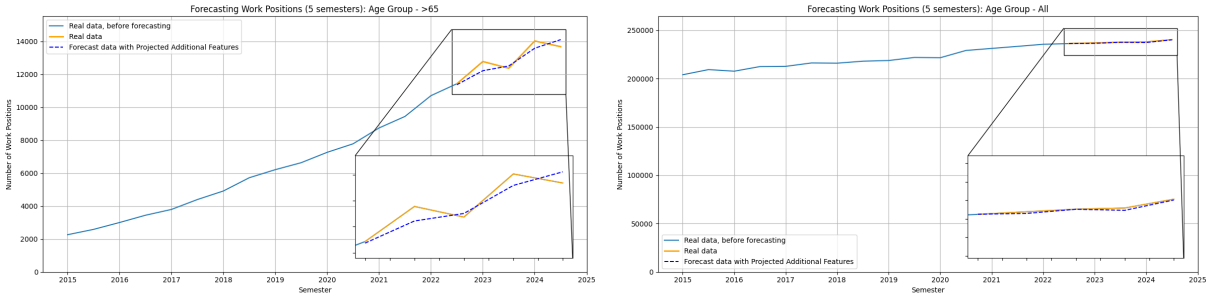
The RMSE and NRMSE values obtained during the training phase, covering the five semesters from the second semester of 2022 to the second semester of 2024, are reported in Table 5.9. Overall, the results exhibit consistently low error levels, with most NRMSE values situated in the thousandths, and only a few extreme age group subsets rising to the hundredths. This indicates a high degree of reliability and stability in the forecasting models when applied to the training horizon.

A recurrent pattern observed in certain scenarios occurs when the model successfully captures the overall trajectory of the data across semesters but underestimates the magnitude of short-term upward or downward fluctuations. This behaviour is particularly evident in the age group above 65 within the Education Ministry, where the gap between actual and forecasted values is illustrated in Figure 5.13a.

Table 5.9. Absolute (rounded to 2 decimals) and Normalised Values (rounded to 3 decimals) in TS Approach with a fixed forecast period of 5 semesters, per Category and Age Group

Category	**Data Type**	<24	25–34	35–44	45–54	55–64	>65	**Total**
Education Ministry	Absolute	68.62	162.52	217.54	186.33	228.23	377.44	361.13
	Normalised	0.045	0.012	0.005	0.002	0.003	0.029	0.002
Early Childhood+Primary/Secondary	Absolute	19.25	53.90	240.54	184.77	205.88	239.23	200.77
	Normalised	0.042	0.014	0.012	0.004	0.004	0.033	0.002
Polytechnic HE Teachers	Absolute	6.65	14.55	43.56	34.49	21.10	13.07	82.11
	Normalised	0.054	0.013	0.015	0.008	0.007	0.037	0.007
University Teachers	Absolute	8.79	30.63	42.97	107.63	44.20	18.17	52.38
	Normalised	0.025	0.014	0.012	0.022	0.009	0.015	0.003

The lowest relative errors are observed in the Education Ministry and Early Childhood Educators and Primary/Secondary Teachers totals, confirming that aggregated forecasts are more stable than age-specific subsets. For the first case, the discrepancy between actual and forecasted values is described in Figure 5.13b.



(a) Age Group: >65 **(b)** Age Group: Total

Figure 5.13. TS Results, for the **Education Ministry**, with a fixed forecast horizon of **5 semesters**

At the opposite end, the highest NRMSE values occur in the youngest and oldest groups, particularly for Polytechnic Higher Education Teachers, as shown in Figure 5.14. These results suggest that forecasts for smaller, more volatile subsets are more prone to error.

In summary, the training phase results, for a fixed forecast horizon of five semesters, demonstrate that this approach delivers strong performance across all categories, with higher effectiveness in forecasting aggregated datasets and slightly reduced capabilities in predicting the extreme age subsets, where data is inherently less stable. These outcomes provide a solid validation of the modelling strategy and establish a solid foundation for extending the forecasts to the testing period.

5.3.1.1. *Best Model for each Category*

The model that achieves the best performance when predicting a fixed horizon of five semesters, for each category and age group, is reported in Table A.3 of Appendix A.

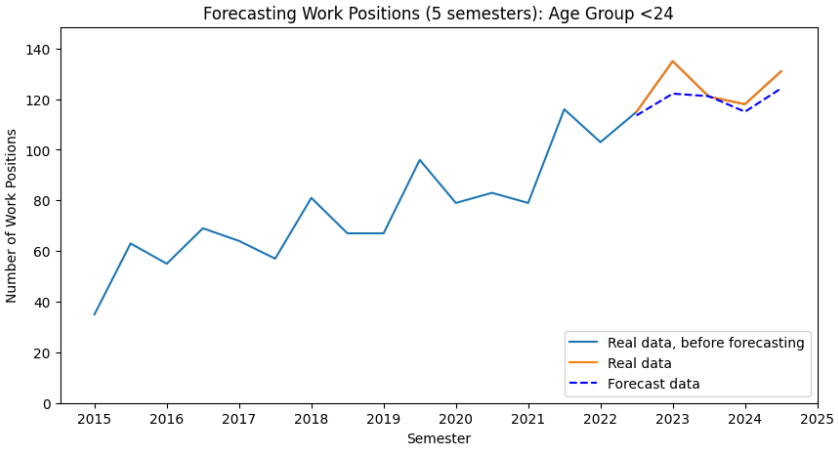


Figure 5.14. TS Results, for **Polytechnic Higher Education Teachers** and divided by Age Group: <24, with a fixed forecast horizon of **5 semesters**

Results confirm the supremacy of the ARIMAX model for age group subsets, as it outperforms all others in 15 out of 24 cases, followed by SARIMAX, in 5 cases, and VAR, in 4 cases. At the complete category level, SARIMAX emerges as the most frequently selected model, appearing 2 times, while ARIMAX and VAR are chosen once each.

These findings highlight the dominance of ARIMAX in predicting higher training–testing percentage subsets, suggesting that seasonality is less prominent in the data. Exponential Smoothing, once again, fails to match the performance of the other models, as it is not selected in this context nor in the broader model evaluation described in Section 5.2.2.4.

When compared to the findings in Section 5.2.2.4, ARIMAX strengthens its position in the age group analysis, increasing from 33.3% to 62.5% of the cases, thereby reinforcing its predominance in this context. In contrast, SARIMAX and VAR show a decline, falling from 31.3% and 35.4% to approximately 20.8% and 16.7%, respectively.

At the full category level, SARIMAX rises from 37.5% to 50%, while ARIMAX drops from 37.5% to 25%. The share of VAR remains stable at 25%.

5.3.2. Future Forecast Predictions

The model and corresponding combination of additional features that achieves the best performance for each category and age subset, as identified in Section 5.3.1, are subsequently employed to generate forecasts for future, unavailable semesters. This testing phase covers the period from the first semester of 2025 to the second semester of 2027. The forecasted number of work positions for each category and age group is reported in Table B.1 of Appendix B.

The forecasts suggest that the Portuguese public education workforce will experience a relatively stable growth trajectory until 2027, with only modest variations across categories and age groups. For the Education Ministry as a whole, the forecasts shown in Figure 5.15 project an increase of nearly 19,000 positions by the end of 2027.

Most subsets follow an upward trend, with some exceptions, such as the 55–64 age group in the Education Ministry, where a gradual decline is observed. At the aggregated

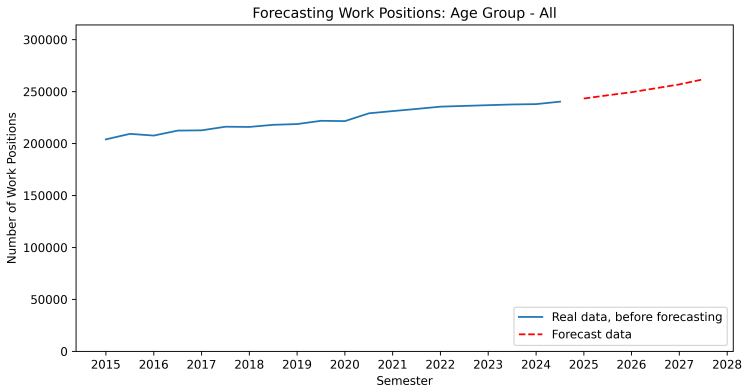


Figure 5.15. Future work positions predictions, for the **Education Ministry**

level, the direct forecasts of the Total dataset indicate a decline in Polytechnic Higher Education Teachers. However, when forecasts are generated separately for each age subset and subsequently summed (Sum of Age Groups), the results reveal a general upward trend.

By contrast, the Early Childhood Educators and Primary/Secondary Teachers dataset shows a more volatile trajectory. While the aggregated forecast indicates a steady decline, the Sum of Age Groups suggests growth from early 2025 until mid-2026, followed by a sharp contraction in June 2027 and a partial recovery by the end of that year.

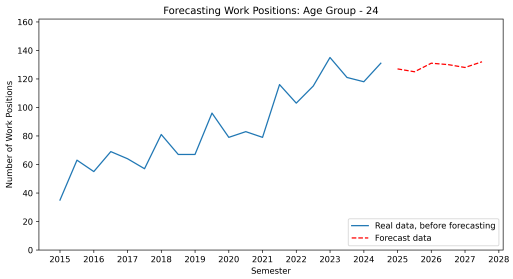
At the extremes of the age distribution, divergent patterns emerge. For Polytechnic Higher Education Teachers under 24 years old, forecasts remain flat at around 130 positions across all semesters, as shown in Figure 5.16a. This stagnation suggests the difficulty of attracting very young professionals to a career path that requires advanced qualifications and prolonged training. Without proactive recruitment or targeted incentives, this cohort is unlikely to expand.

Conversely, the age group above 65 in the same category follows a downward trajectory, described in Figure 5.16b. In other categories, however, the >65 cohort shows small but steady increases, reinforcing the ageing of the teaching workforce. This pattern aligns with the ageing workforce issues described in [3] and further examined in Section 3.2.1, although the proportion of teachers aged 50 and above is slightly lower than the one introduced in the study.

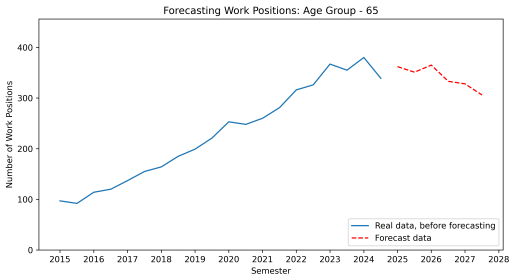
While reliance on older teachers may temporarily mitigate workforce shortages, the continued ageing of the teaching workforce poses long-term risks, as eventual retirements could create critical gaps and threaten the sustainability of teacher supply.

5.3.2.1. *Comparison of Predicted Work Positions between Total Forecast and Sum of Individual Age Groups Forecast*

The comparison between the total forecast and the sum of the individual age group forecasts provides insight into the internal consistency of the modelling approach. Table 5.10 reports the MAE and MPE values, which quantify the absolute and relative differences



(a) Age Group: <24



(b) Age Group: >65

Figure 5.16. Future work positions predictions, for **Polytechnic Higher Education Teachers**

between the aggregated age group forecasts and the direct total forecasts. While the deviations remain modest in absolute terms, they are noticeably higher than those observed during the model assessment phase, where all relative errors remained below 0.5%.

Table 5.10. MAE and MPE between Summed Age Group future forecasts and Total future forecasts (using Projected Additional Features, rounded to 2 decimals)

Model	**MAE**	**MPE (%)**
Education Ministry	7,428.19	2.902
Early Childhood + Primary/Secondary Teachers	2,501.48	1.947
Polytechnic HE Teachers	935.72	8.315
University Teachers	480.15	2.531

For the Education Ministry, the MPE of 2.90% indicates that summing individual age group forecasts slightly overestimates the total workforce, representing a deterioration relative to the high performance achieved in the training–testing evaluation. Early Childhood Educators and Primary/Secondary and University Teachers exhibit similar behaviour, suggesting that errors accumulate when forecasting further into the future. The Polytechnic Higher Education Teachers present the largest relative deviation, with an MPE of 8.32%, highlighting that in smaller workforce categories, even moderate absolute differences can result in substantial relative errors.

Taken together, these results show a mixed picture for the Portuguese teaching workforce. Forecasts suggest relative stability in the short term, but the ongoing ageing of teachers and the low number of younger professionals entering the field raise long-term concerns. Relying on older teachers may temporarily fill gaps, but it could create generational imbalances that threaten the sustainability of the education system. These findings highlight the importance of attracting younger teachers, supporting earlier career entry and maintaining a balanced distribution of teachers across age groups.

The results largely confirm the patterns observed in Section 3.2, as across all categories the workforce is concentrated in intermediate age groups, while the youngest (<24 and 25–34) and oldest (>65) subsets consistently show the lowest numbers.

5.4. Future Forecast for Contract Type and Region

To estimate future work positions according to their contract type, the same forecasting procedure described in Section 5.3 is implemented. This involves applying the best-performing TS models for each category, together with the corresponding projected additional features, thereby extending the approach used for age group forecasts to the contract-level analysis.

As this dataset represents the number of contracts associated with the Early Childhood Educators and Primary/Secondary Teachers position, the additional feature of Registered Students, used in both the training and testing phases, corresponds exclusively to the number of students enrolled in the relevant Early Childhood Education and Primary/Secondary levels.

In this section, the insular regions of Madeira and Azores are also incorporated, as they form part of the NUTS III regions, consistent with the rest of the dataset.

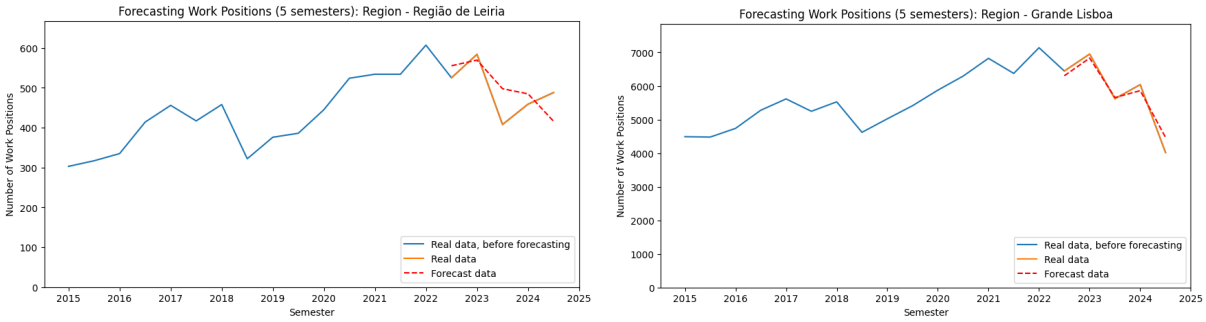
5.4.1. Future Forecast Model Performance

Tables B.2 and B.3 of Appendix B present the absolute and normalised forecast RMSE values per contract type and NUTS III region, obtained using the TS approach with a fixed forecast horizon of five semesters. The normalised values offer a clearer perspective on relative performance across regions and contract types, enabling comparisons independent of scale. Overall, the NRMSE values remain low in most regions, indicating strong predictive performance, particularly for Permanent Contracts.

For Fixed-Term Contracts, NRMSE values are typically in the hundredths or low tenths, reflecting the inherent difficulty in forecasting contracts that are less stable and influenced by independent events or emergent needs. The differences between actual and forecasted work positions for the regions of *Região de Leiria* and *Grande Lisboa* are illustrated in Figures 5.17a and 5.17b, respectively. In the case of *Região de Leiria*, the forecasts fail to fully capture the trend of the period after 2023, whereas in *Grande Lisboa*, the forecasted values closely follow the observed trajectory, demonstrating strong predictive alignment.

For Permanent Contracts, NRMSE values are significantly lower, typically in the thousandths and occasionally in the low hundredths. This reflects the uniformity and stability of these contracts, which generally follow a consistent trend across semesters. Unlike Fixed-Term Contracts, Permanent Contracts have no predefined end date, resulting in minimal fluctuations over time.

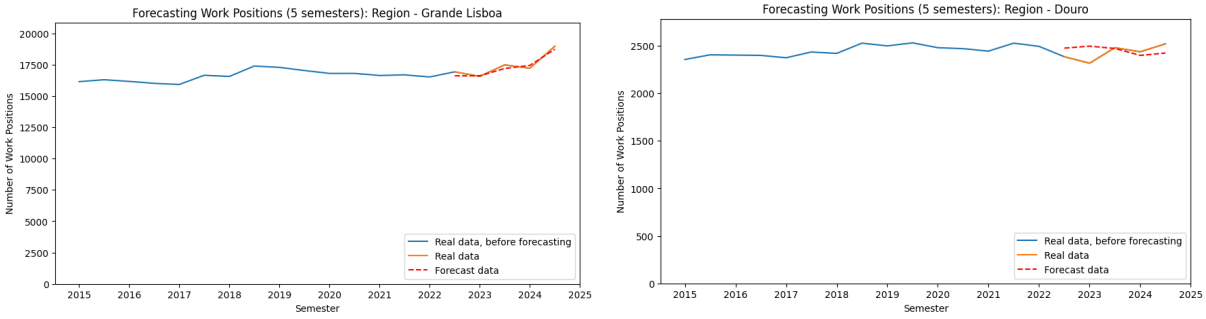
In the case of *Grande Lisboa*, the forecasted number of Permanent Contracts closely follows the trajectory of the actual values, as illustrated in Figure 5.18a, where the predicted pattern aligns almost perfectly with the observed data. In contrast, for regions such



(a) Region: *Região de Leiria* **(b)** Region: *Grande Lisboa*

Figure 5.17. TS Results, for **Fixed-Term Contracts**, with a fixed forecast horizon of **5 semesters**

as *Douro*, the model occasionally struggles to capture the precise direction of fluctuations, at times exhibiting trends opposite to those observed in the actual data. Nevertheless, the projected values remain generally close to the actual values, particularly during the second semester of 2023, as shown in Figure 5.18b.

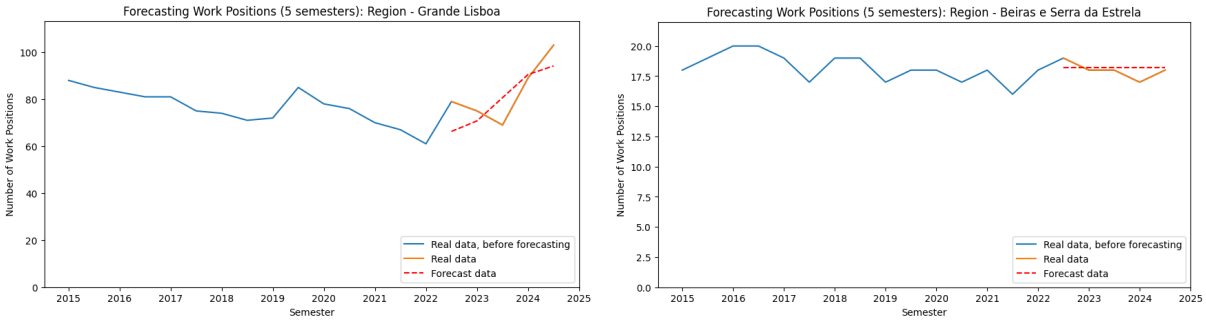


(a) Region: *Grande Lisboa* **(b)** Region: *Douro*

Figure 5.18. TS Results, for **Permanent Contracts**, with a fixed forecast horizon of **5 semesters**

Finally, Commission Service or Political Office/Mandate Contracts, designated as 'Commission Service' in the tables, generally exhibit relative errors in the same order as Fixed-Term Contracts, typically ranging from the thousandths to low hundredths. This is attributable to their small scale and greater variability. Notably, the insular regions of Madeira and Azores (*Região Autónoma da Madeira* and *Região Autónoma dos Açores*, respectively) do not include any Commission Service or Political Office/Mandate Contracts, explaining the absence of values for these categories.

In *Grande Lisboa*, the model captures the overall upward trajectory but fails to reflect the sharp decline in work positions at the end of 2023, resulting in limited accuracy, as shown in Figure 5.19a. In the case of *Beiras e Serra da Estrela*, the model forecasts a constant number of positions across all semesters, as illustrated in Figure 5.19b. For other regions, such as *Alentejo Central* (Figure 5.20), the model follows the overall trend of the actual values, despite some discrepancies, converging closer to the observed values after the end of 2023.



(a) Region: *Grande Lisboa* **(b)** Region: *Beiras e Serra da Estrela*

Figure 5.19. TS Results, for **Commission Service or Political Office/Mandate Contracts** in *Beiras e Serra da Estrela*, with a fixed forecast horizon of **5 semesters**

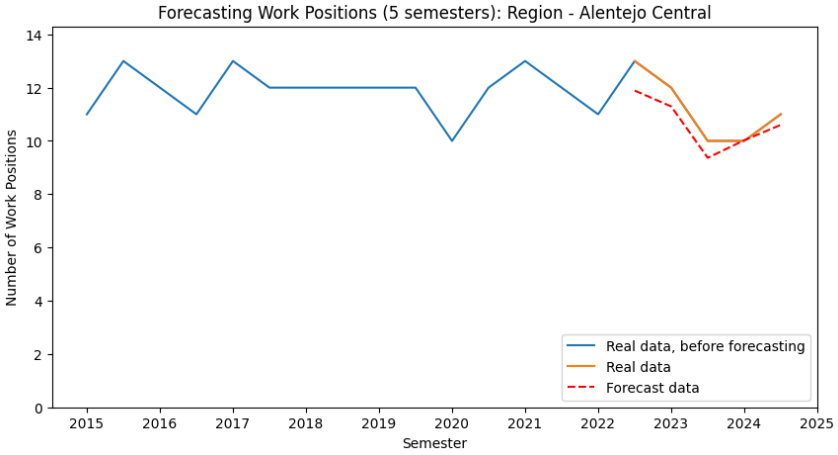


Figure 5.20. TS Results, for **Commission Service or Political Office/Mandate Contracts** in *Alentejo Central*, with a fixed forecast horizon of **5 semesters**

This pattern highlights that the models perform most reliably for the more numerous and stable Permanent Contracts, which exhibit consistent trends over time. In contrast, contract types that are less frequent, more variable or subject to external and situational factors, such as Fixed-Term or Commission Service or Political Office/Mandate Contracts, pose greater challenges for accurate forecasting, resulting in higher relative errors and reduced predictive reliability.

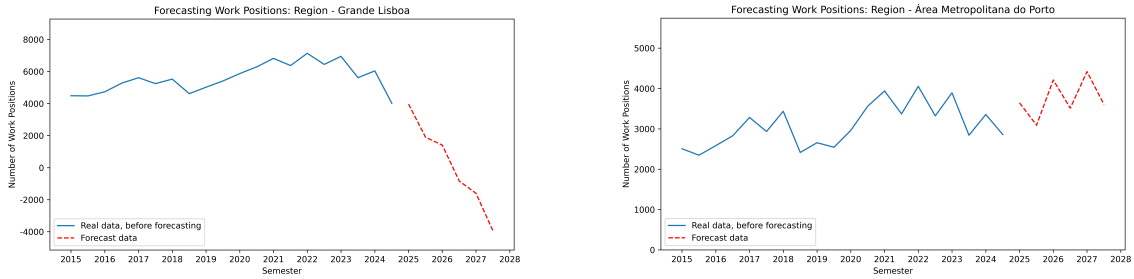
5.4.2. Future Forecast Predictions

The forecasts of work positions per contract type, obtained using the TS approach with a fixed horizon of six semesters (2025–2027), are reported in Tables B.7 to B.10 of Appendix B. The aggregated totals are also shown in the plots, with negative values rounded to zero in order to avoid accounting for the unrealistic case of a negative number of positions under certain contract types.

Overall, the projections suggest that Permanent Contracts will continue to dominate the workforce across most regions, while Fixed-Term and Commission Service or Political Office/Mandate Contracts exhibit higher variability and less predictable patterns.

5.4.2.1. *Future Predictions for Fixed-Term Contracts*

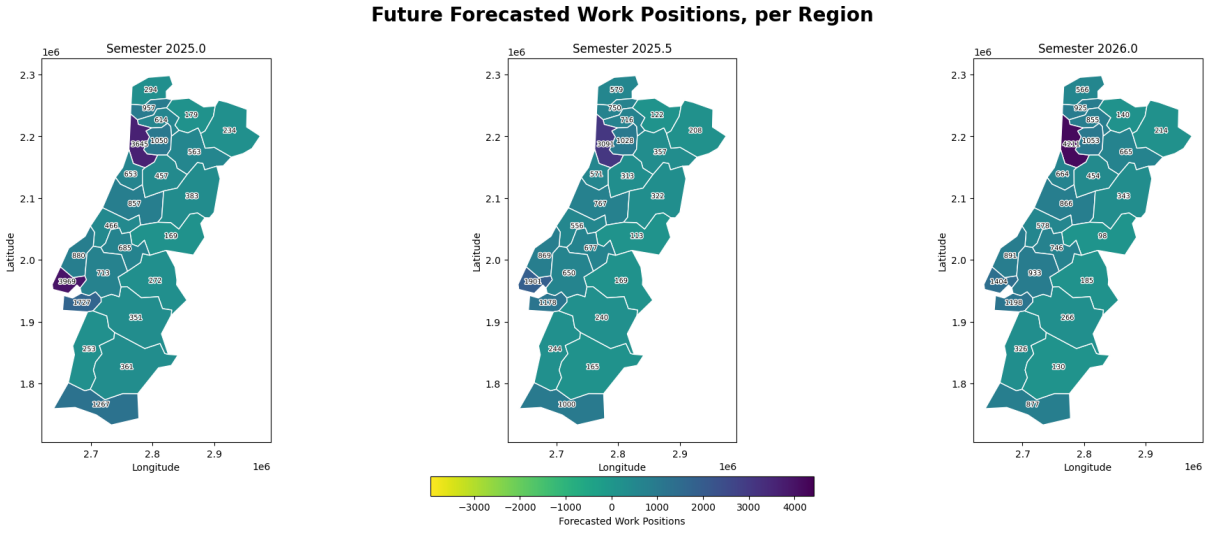
For Fixed-Term Contracts, several regions exhibit irregular trajectories, with some projections even producing negative values, which are unrealistic. This outcome underscores the inherent difficulty of forecasting such irregular data. In *Grande Lisboa*, for example, negative values appear in the second half of 2026 and in 2027, as shown in Figure 5.21a. This suggests potential model instability, compounded by the unpredictability of historical data and possible limitations in model calibration during the training phase. By contrast, other regions, such as *Área Metropolitana do Porto*, display more consistent forecasts that align more closely with past trends, as illustrated in Figure 5.21b.



(a) Region: *Grande Lisboa* **(b)** Region: *Área Metropolitana do Porto*

Figure 5.21. Future work positions predictions, for **Fixed-Term Contracts**

To provide a clearer assessment of the forecasted workforce positions across regions, Figure 5.22 presents the spatial forecast distribution of predicted positions regarding Fixed-Term Contracts, for each NUTS III region over the forecasting horizon. This spatial perspective highlights regional asymmetries and complements the future forecast analyses, enabling a more comprehensive understanding of workforce dynamics across Portugal.



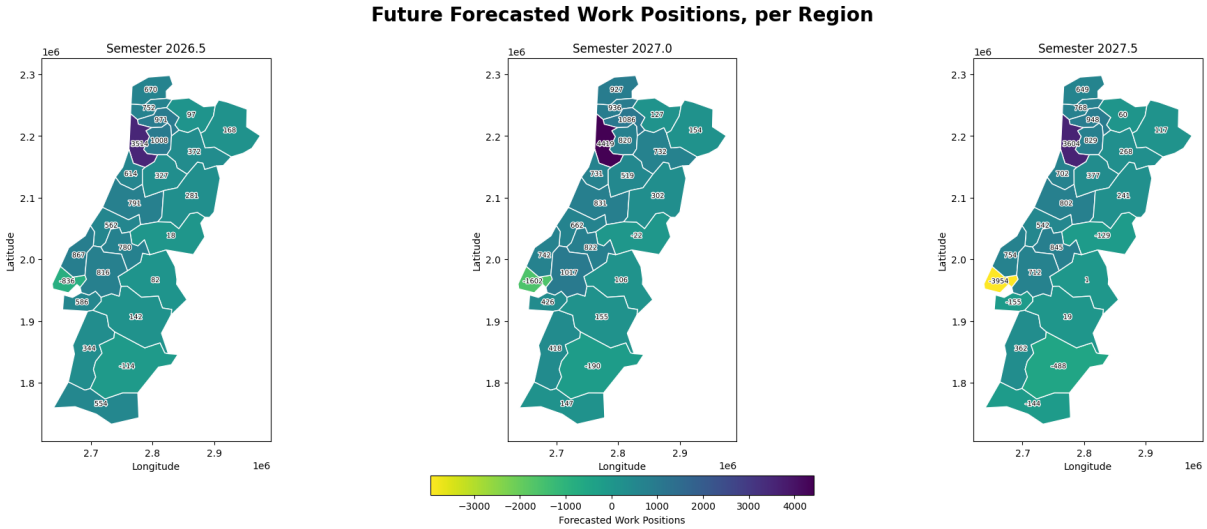


Figure 5.22. Evolution of forecasted **Fixed-Term Contracts** per NUTS III Regions

5.4.2.2. *Future Predictions for Permanent Contracts*

Permanent Contracts exhibit a notably stable and predictable trajectory, reflecting their long-term and continuous nature. For instance, in *Grande Lisboa*, the projected number of Permanent Contracts steadily rises from 19,569 in the second semester of 2025 to 31,425 by late 2027, indicating sustained workforce growth, as illustrated in Figure 5.23.

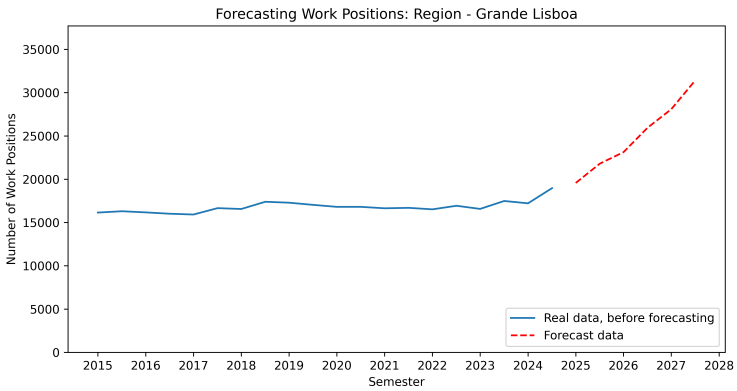
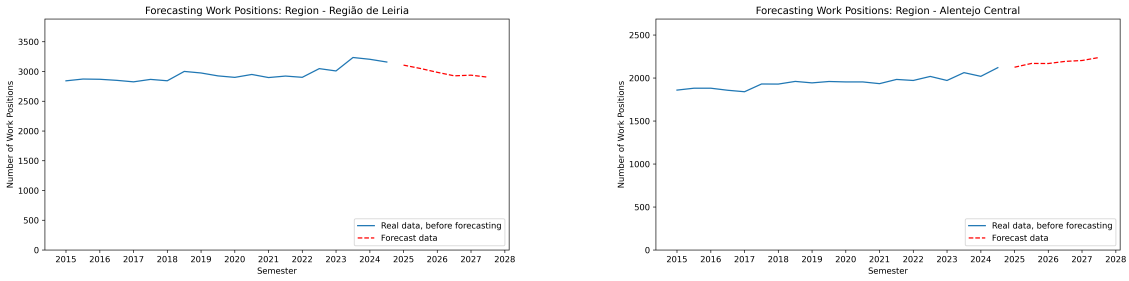


Figure 5.23. Future work positions predictions, for **Permanent Contracts**, in *Grande Lisboa*

In contrast, regions such as *Região de Leiria* display a modest decline in Permanent Contract positions over the same period, as shown in Figure 5.24a, highlighting regional variations in workforce dynamics. Smaller regions, including *Alentejo Central* (Figure 5.24b), exhibit consistent growth, underscoring the resilience and stability of Permanent Contracts even in less populous areas.

Figure 5.25 illustrates the spatial distribution of forecasted Permanent Contract positions across all NUTS III regions throughout the projection horizon.



(a) Region: *Região de Leiria* **(b)** Region: *Alentejo Central*

Figure 5.24. Future work positions predictions, for **Permanent Contracts**

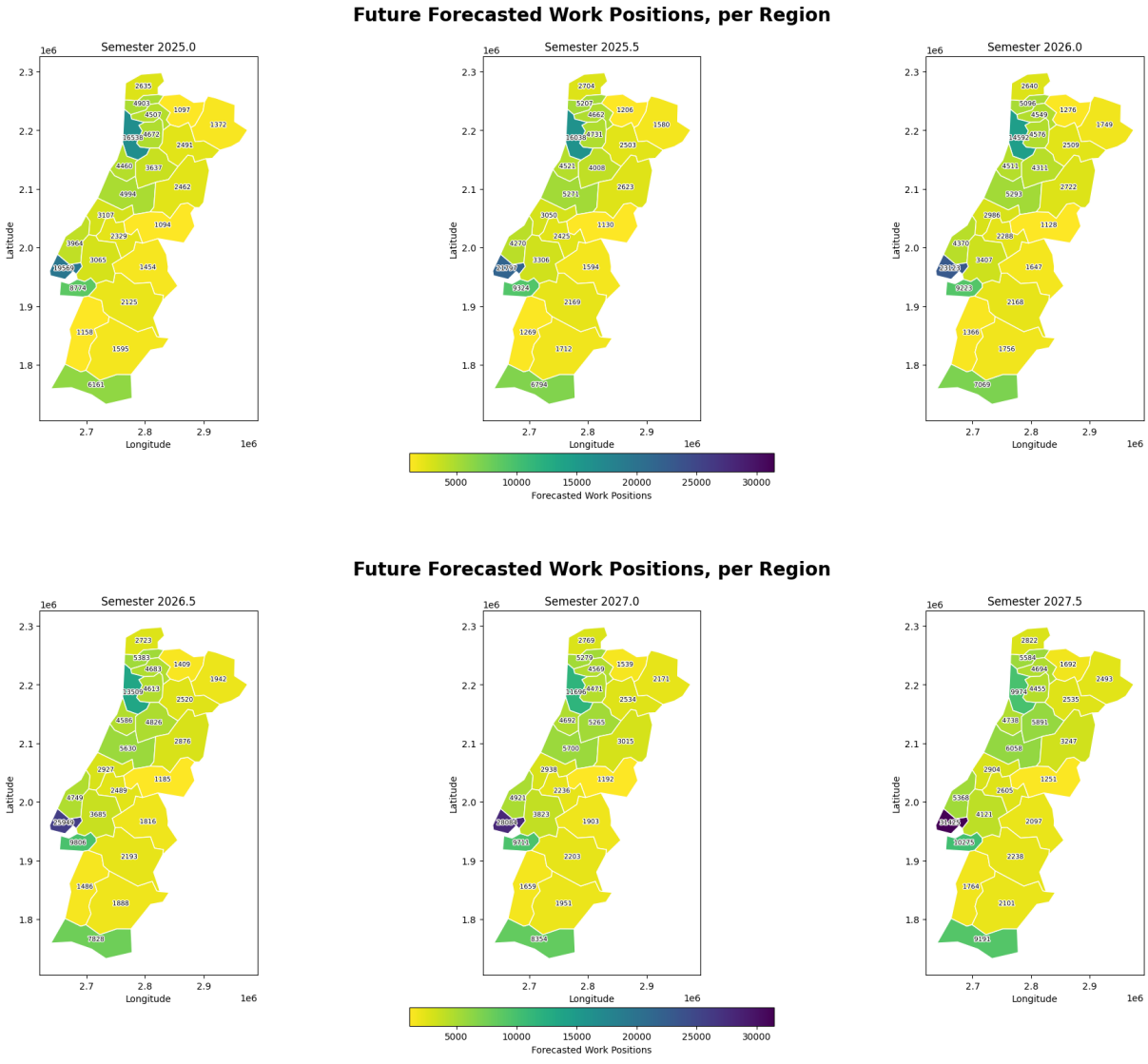


Figure 5.25. Evolution of forecasted **Permanent Contracts** per NUTS III Regions

5.4.2.3. *Future Predictions for Commission Service or Political Office/Mandate Contracts*

Commission Service or Political Office/Mandate Contracts remain limited in number and are intrinsically challenging to forecast due to their irregular and unpredictable nature.

Across most regions, the projected number of positions remains largely stable, with minimal fluctuations. For example, in *Beiras e Serra da Estrela*, all forecasted semesters consistently indicate 18 positions throughout the projection period, as illustrated in Figure 5.26, reflecting a small but stable workforce in this contract category.

In *Grande Lisboa*, the forecast suggests a gradual increase, from 112 positions in June 2025 to 167 by December 2027, as shown in Figure 5.27a, indicating localised growth in political or mandate-related roles. Conversely, the specific region of *Médio Tejo* exhibits a sharp decline, from 7 positions at the beginning of the forecast period to -14 at the end, as depicted in Figure 5.27b, exhibiting an impracticable pattern.

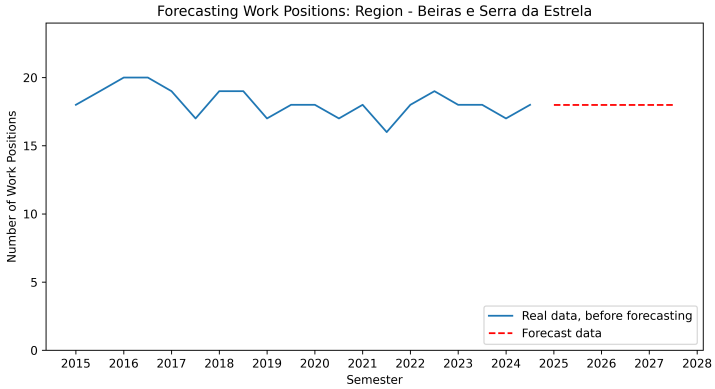
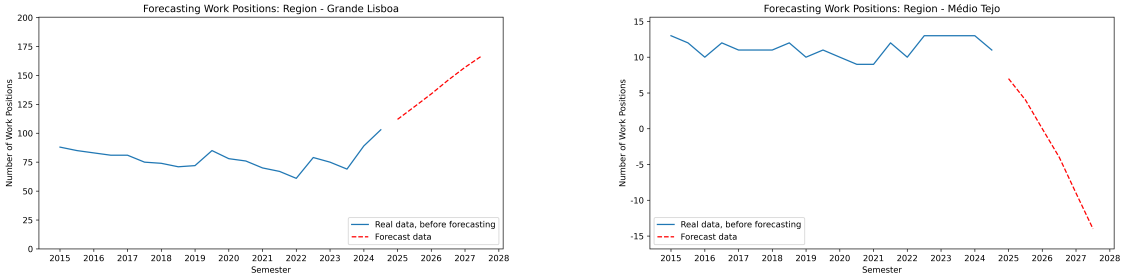


Figure 5.26. Future work positions predictions, for **Commission Service or Political Office/Mandate Contracts**, in *Beiras e Serra da Estrela*



(a) Region: *Grande Lisboa* **(b)** Region: *Médio Tejo*

Figure 5.27. Future work positions predictions, for **Commission Service or Political Office/Mandate Contracts**

Figure 5.28 presents the regional forecast patterns of Commission Service or Political Office/Mandate Contracts, mapping their distribution across all NUTS III regions over the forecasting period.

Overall, these results suggest that Permanent Contracts will continue to make up the majority of the Portuguese teaching workforce, showing a stable and predictable increase across most regions. In contrast, Fixed-Term and Commission Service or Political Office/Mandate Contracts are less predictable, with irregular trends and regional differences

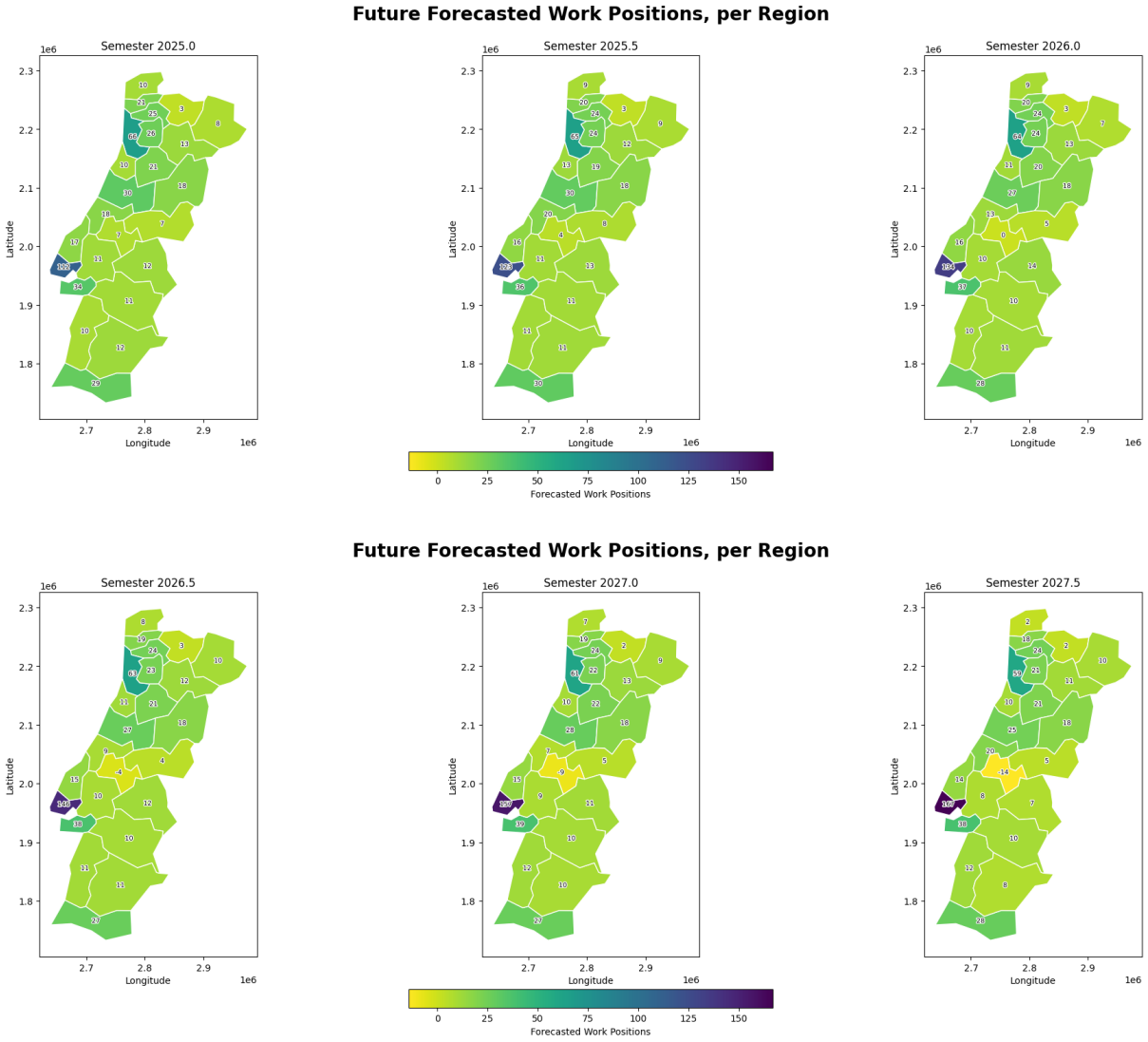


Figure 5.28. Evolution of forecasted **Commission Service or Political Office/Mandate Contracts** per NUTS III Regions

and, in some cases, unrealistic projections caused by data volatility or model limitations. The results also support the analysis from Section 3.2.4, showing that metropolitan areas like Lisbon and Porto hold the largest share of teachers, with a clear contrast between their workforce levels and those of smaller regions.

A critical challenge in this phase of the study concerns the change in data sources for the additional features of Public Debt and Unemployment Rate, since projections for these variables are not available across the entire forecasting horizon. To address this gap, alternative sources are adopted for the missing semesters.

However, this adjustment has a noticeable impact on the results for 2027, where forecasts in models using these features tend to fall below zero or exhibit sharp declines. This behaviour contrasts sharply with the forecasts of models that exclude these features, which generally display more plausible and consistent trajectories. The features used in each combination of Contract Type and Region models are presented in Tables B.4 to B.6. These tables illustrate that when these specific features are included, the forecasts

tend to be less reliable, whereas models that either exclude them or combine them with other features yield more trustworthy predictions.

While the impact of this source change is not particularly evident in the earlier analyses, its influence becomes pronounced in this case, clearly differentiating the results depending on whether or not the affected features are included.

All the plots regarding the forecast of future semesters are available in Appendix C.

CHAPTER 6

Conclusions

This study addresses the lack of research on Portuguese public employment by proposing a comprehensive framework for analysing and forecasting workforce positions in the education sector. The work is structured around three research questions:

- ***RQ1.*** *Can tendencies and patterns be identified in public employment data, in the teaching sector, in the past few years?*
- ***RQ2.*** *Can future public employment be forecast through the analysis and investigation of related past processes?*
- ***RQ3.*** *How do TS and ML methods contribute to improving the predictive accuracy in forecasting public employment?*

To address **RQ1**, an exploratory dataset analysis is carried out, including correlations among variables, long-term trends in teaching positions and workforce distributions by gender, age group and contract type. Regional disparities across NUTS III are also examined, complemented by per-capita and student-based indicators. The analysis confirms structural patterns in the workforce, especially a persistent ageing trend and strong territorial imbalances.

To answer **RQ2**, two baseline simulations are developed: one relying solely on the ageing process and another introducing a delta variation to capture variability. While these simulations produce reasonable outputs in some cases, their reliance on perturbations of observed data limits their robustness, highlighting the need for more advanced forecasting approaches.

RQ3 is addressed by implementing both TS and ML forecasting techniques. TS models such as ARIMA, SARIMAX and VAR are compared against ML methodologies like MLP and Gradient Boosting, using different training-testing splits. External explanatory variables, like Minimum Wage and GDP, are also incorporated to enhance predictive performance.

Results demonstrate that TS methods consistently outperform both baseline and ML approaches, delivering forecasts that are accurate, stable and well-suited to this type of data. While ML techniques show potential for capturing complex and non-linear patterns, their high variability and sensitivity limit their reliability. The inclusion of external features significantly improves results, underscoring the importance of social and economic factors in such forecasts.

Projections confirm a persistent ageing trend in the education workforce, with older groups gaining dominance while younger cohorts remain under-represented. Permanent

contracts appear stable, but fixed-term and commission service or political office/mandate positions show higher volatility. Regional disparities are clear, with Lisbon and Porto concentrating the majority of teachers. Together, these findings expose structural challenges of generational renewal and territorial imbalance, with the risk of future shortages as retirements accelerate.

Overall, this work achieves its objectives and answers the proposed research questions, offering both methodological insights and practical guidance. It confirms the superiority of TS-based forecasting for this context, while contributing to the scarce literature on public employment forecasting in Portugal.

An article presenting this research, including its methodology and results, was submitted to *Science and Public Policy*, a leading international journal on public policies for science, technology and innovation, published by Oxford University Press.

6.1. Limitations

Several constraints related to the data and methodologies used can be identified as potential limitations of this study. These limitations arise primarily from the nature of the data sources, their coverage and the need for adjustments to ensure comparability across datasets.

Potential inconsistencies across datasets

Although all real data is obtained from the same statistical portal [31], the number of registered students is reported in two separate datasets: one for Early Childhood Educators and Primary/Secondary Teachers, and another for Polytechnic Higher Education and University Teachers. While their structures are consistent, the fact that they are provided as separate files may imply subtle differences in data collection or processing methodologies. Although not necessarily problematic, this introduces the possibility of minor inconsistencies when aggregating values, particularly in the case of the Education Ministry dataset, which combines both.

Temporal resolution mismatch and interpolation

For the actual past values of additional features, there is a temporal mismatch, as the main dataset is reported on a semester basis, whereas the complementary datasets are only available annually. To ensure consistency, interpolation is applied to estimate values for the first semester of each year. This method preserves overall seasonal patterns and allows alignment with the primary dataset. However, it may introduce distortions and represent a potential source of inaccuracy in model assessments.

Divergence of projection sources

Projections for explanatory variables in the testing phase are derived from multiple institutions, as Unemployment Rate and Public Debt are extracted from OECD and GDP and GFCF from *Banco de Portugal*. In contrast, Minimum Wage and student registrations are based on less rigorous methods, including government announcements in the case of

the Minimum Wage and AGR-based extrapolations for student registrations. These simplified approaches lack a strong and verified foundation, thereby introducing additional uncertainty into the long-term forecasting of workforce demand.

Limitations of projection horizons

OECD only have available projections up to 2026. Beyond this horizon, data from *Banco de Portugal* is used for 2027. This mixing of sources, with potentially divergent assumptions and methodologies, may influence the results. As highlighted in Section 5.4.2.2, this limitation is particularly relevant for categories or regions where Unemployment Rate and Public Debt play a significant role in the best-performing models, as the different projection bases likely affect the forecasts. An important future improvement would be to use consistent projections for all years, including the entire testing period.

Data splitting constraints

Another potential limitation concerns the way data is divided for model evaluation and future forecasting. In the initial model assessment phase, data is split by percentages. However, due to the limited number of semesters available, exact percentage splits cannot always be achieved. The division approximates the intended proportions, but minor deviations are inevitable. One possible way to address this issue would be to perform model evaluation based on the number of semesters rather than percentage splits.

This issue extends to the future forecasting phase: for example, using five semesters out of twenty for training does not correspond to the same percentage as using six semesters out of twenty-six for testing. The proportions are close but not identical, differing by roughly 0.2 when rounded. Although the discrepancy is small and unlikely to substantially affect results, it represents a practical constraint of working with limited semester-level data.

Computational restraints of ML models

The complexities of some ML models, especially MLP, represent a notable constraint in this study, as this model demands substantial computational resources and time, which limit the number of experiments and robustness checks performed. Training multiple models for each category and age group under different forecasting horizons amplifies these computational requirements.

A practical improvement would be to explore more efficient training algorithms and model simplifications, to mitigate these limitations and allow more extensive evaluations.

6.2. Future Work

This section outlines potential directions for improving and extending this study, including methodological refinements, incorporation of additional data sources and exploration of new analytical perspectives.

Incorporation of other external variables

A potential path for future work is the inclusion of other exogenous variables in the study and the models to complement and enhance the current feature set. The use of external variables has consistently improved model performance, as reflected by lower RMSE values compared to models without such features. Incorporating further relevant variables could provide additional benefits, potentially leading to more reliable and robust forecasts.

Integration of additional models

The application of other models and methodologies reveals a promising direction for future work. The inclusion of hybrid models, which combine TS and ML techniques, could enhance predictive performance by leveraging the strengths of both paradigms.

Moreover, the application of more complex neural networks and deep learning architectures appears well-suited to the problem at hand, offering potential improvements in capturing non-linear patterns and intricate temporal dependencies. Investigating these advanced models could provide deeper insights and further refine the forecasting framework established in this study.

Influence of external factors

In terms of external factors, such as education policies, curricular reforms, government changes and shifts in the parties in power, each often bringing different priorities and measures for the sector, this study does not account for their potential effects.

These factors can significantly influence workforce demand and contract stability, sometimes even more strongly than demographic or economic variables. A meaningful improvement for future research would be the inclusion of such elements and the analysis of their direct and indirect impacts on forecasting results, as they represent structural forces shaping the long-term evolution of the education system.

Creation of alternative scenarios

To better account for potential changes in the education sector, the creation and analysis of alternative scenarios could be a valuable extension of this study. Examples include variations in the number of registered students, changes in retirement age or adjustments in class sizes or staffing rules.

Incorporating these hypothetical scenarios would allow for a more comprehensive understanding of how different factors might influence workforce projections, providing decision-makers with a more robust tool for planning under uncertainty.

Extension to other regions and sectors

Following the assessment conducted in this study, which focuses on the Portuguese public education sector and its workforce, a similar evaluation can be carried out for the insular regions of Madeira and Azores, which are not included in the current principal analysis.

Additionally, extending this methodology to other sectors of Portuguese public employment, such as healthcare, justice or social services, can provide valuable insights into

workforce dynamics across different public domains and evaluate the applicability of the proposed models.

Applications for public education decision-making

Finally, the analysis and results of this and similar studies can be transformed into practical tools to assist organisations such as schools, government institutions and educational management bodies in Portugal, allowing them to extract valuable information and support their decision-making. These tools can provide insights that have a tangible impact on workforce planning, policy choices and management of the public education sector.

[This page is intentionally left blank.]

References

[1] Direção-Geral de Estatísticas da Educação e Ciência (DGEEC), *Perfil do Aluno 2022/2023*. Lisboa, Portugal, Sep. 2024, ISBN: 978-972-614-840-1. [Online]. Available: http://www.dgeec.medu.pt.

[2] Euronews. "Portugal faces teacher shortage, leaving thousands without educators." Accessed: 10 September 2025. (Sep. 2024), [Online]. Available: https://www.euronews.com/2024/09/12/portugal-faces-teacher-shortage-leaving-thousands-without-educators.

[3] OECD, *TALIS 2018 Results (Volume I): Teachers and School Leaders as Lifelong Learners*. Paris: OECD Publishing, Jun. 2019, ISBN: 9789264752566. DOI: 10.1787/1d0bc92a-en. [Online]. Available: https://doi.org/10.1787/1d0bc92a-en.

[4] OECD, *TALIS 2013 Results: An International Perspective on Teaching and Learning*. Paris: OECD Publishing, Jun. 2014, ISBN: 9789264211339. DOI: 10.1787/9789264196261-en. [Online]. Available: https://doi.org/10.1787/9789264196261-en.

[5] DIOEP - Departamento de Informação da Organização do Estado e do Emprego Público, "Siep - síntese estatística do emprego público," Feb. 2025, ISSN: 2182-7311. [Online]. Available: https://www.dgaep.gov.pt/index.cfm?OBJID=ECA5D4CB-42B8-4692-A96C-8AAD63010A54.

[6] D.-G. da Administração e do Emprego Público (DGAEP), *Boletim estatístico do emprego público (boep)*, Jun. 2025. [Online]. Available: https://www.dgaep.gov.pt/index.cfm?OBJID=C0F56E62-5381-4271-B010-37ECE5B31017.

[7] M. J. Page, D. Moher, P. M. Bossuyt, I. Boutron, T. Hoffmann, C. Mulrow, L. Shamseer, J. Tetzlaff, and et al., "Prisma 2020 explanation and elaboration: Updated guidance and exemplars for reporting systematic reviews," *BMJ*, vol. 372, 2021. [Online]. Available: https://www.bmj.com/content/372/bmj.n160.

[8] L. J. de Sousa, M. L. Simões, J. P. Martins, L. Sanhudo, and J. M. da Costa, "Statistical descriptive analysis of portuguese public procurement data from 2015 to 2022," *CivilEng*, vol. 4, pp. 808–826, 3 Sep. 2023, ISSN: 26734109. DOI: 10.3390/civileng4030045.

[9] L. J. D. Sousa and J. P. Martins, "Predicting construction project compliance with machine learning model: Case study using portuguese procurement data," *Data in Brief*, vol. 48, pp. 285–302, 13 2015. DOI: 10.1016/j.dib.2023.

[10] Eurostat. "Healthcare personnel statistics - physicians." Accessed: 01 March 2025. (2024), [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Healthcare_personnel_statistics_-_physicians.

[11] A. Sarmento, B. Gomes, J. I. G. Fragata, M. T. Veríssimo, M. Oliveira, N. Sousa, P. Santana, T. Grilo, and T. Gonçalves, *Relatório do grupo de trabalho para a avaliação das necessidades formativas em medicina*, 2024.

[12] L. C. Nunes, A. B. Reis, P. Freitas, M. Nunes, and J. M. Gabriel, *Estudo de diagnóstico de necessidades docentes de 2021 a 2030*. 2021, ISBN: 978-972-614-744-2. [Online]. Available: http://www.dgeec.mec.pt.

[13] W.-C. Hong, S.-Y. Wei, and Y.-F. Chen, "A comparative test of two employee turnover prediction models," *The International Journal of Management*, vol. 24, p. 216, Dec. 2006. [Online]. Available: https://api.semanticscholar.org/CorpusID:152861469.

[14] D. Coleman, "Does europe need immigrants? population and work force projections," *International Migration Review*, vol. 26, no. 2, pp. 413–461, 1992, Cited by: 51. DOI: 10.2307/2547066. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0026879405&doi=10.2307%2f2547066&partnerID=40&md5=d489fc5d7986a66be87fe574be09c6f7.

[15] P. I. Buerhaus, D. O. Staiger, and D. I. Auerbach, "Implications of an aging registered nurse workforce," *JAMA*, vol. 283, no. 22, pp. 2948–2954, 2000, Cited by: 448. DOI: 10.1001/jama.283.22.2948. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0034647326&doi=10.1001%2fjama.283.22.2948&partnerID=40&md5=917ca909747acf88ccfdd5f044fb74d1.

[16] J. L. Zimbelman, S. P. Juraschek, X. Zhang, and V. W.-H. Lin, "Physical therapy workforce in the united states: Forecasting nationwide shortages," *PM and R*, vol. 2, no. 11, pp. 1021–1029, 2010, Cited by: 65. DOI: 10.1016/j.pmrj.2010.06.015. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-78549257761&doi=10.1016%2fj.pmrj.2010.06.015&partnerID=40&md5=ff226f529ed8f1ad46d91c99eb5899b2.

[17] M. D. Landry, L. M. Hack, E. Coulson, J. Freburger, M. P. Johnson, R. Katz, J. Kerwin, M. H. Smith, H. C. "Bud" Wessman, D. G. Venskus, P. L. Sinnott, and M. Goldstein, "Workforce projections 2010–2020: Annual supply and demand forecasting models for physical therapists across the united states," *Physical Therapy*, vol. 96, no. 1, pp. 71–80, 2016, Cited by: 52; All Open Access, Bronze Open Access, Green Open Access. DOI: 10.2522/ptj.20150010. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84953205962&doi=10.2522%2fptj.20150010&partnerID=40&md5=dbc50dc3e8692d748d8be43ddda54516.

[18] isee systems, *Stella*, version 9.1, [Computer Program], Lebanon, NH, 2010. [Online]. Available: https://www.iseesystems.com.

[19] T. Dall, R. Reynolds, R. Chakrabarti, C. Ruttinger, P. Zarek, and O. Parker, *The Complexities of Physician Supply and Demand: Projections from 2021 to 2036*. Mar. 2024.

[20] S. F. Witt, H. Song, and S. Wanhill, "Forecasting tourism-generated employment: The case of denmark," *Tourism Economics*, vol. 10, no. 2, pp. 167–176, 2004, Cited by: 54; All Open Access, Green Open Access. DOI: 10.5367/000000004323142407. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-3042820606&doi=10.5367%2f000000004323142407&partnerID=40&md5=a1f026e0774b708f50f0dc9ed26d4d33.

[21] J. P. LeSage, "Forecasting metropolitan employment using an export-base error-correction model," *Journal of Regional Science*, vol. 30, no. 3, pp. 307–323, 1990, Cited by: 41. DOI: 10.1111/j.1467-9787.1990.tb00102.x. [Online]. Available: https://www.scopus.com/pages/publications/0025622430?inward.

[22] S. Desiere, K. Langenbucher, and L. Struyven, *Statistical profiling in public employment services*, Feb. 2019. DOI: 10.1787/b5e5f16e-en. [Online]. Available: https://www.oecd.org/en/publications/statistical-profiling-in-public-employment-services_b5e5f16e-en.html.

[23] K. Rabuzin and N. Modrušan, "Prediction of public procurement corruption indices using machine learning methods," Cited by: 23; All Open Access, Hybrid Gold Open Access, vol. 3, 2019, pp. 333–340. DOI: 10.5220/0008353603330340. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85074160553&doi=10.5220%2f0008353603330340&partnerID=40&md5=c4c312b6c091306d81b81bc5fefe73c4.

[24] Y. Wei, X. Rao, Y. Fu, L. Song, H. Chen, and J. Li, "Machine learning prediction model based on enhanced bat algorithm and support vector machine for slow employment prediction," *PLoS ONE*, vol. 18, 11 November Nov. 2023, ISSN: 19326203. DOI: 10.1371/journal.pone.0294114.

[25] J. Wikström, "Employment forecasting using data from the swedish public employment service," Ph.D. dissertation, 2018. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-239174.

[26] S. S. W. Fatima and A. Rahimi, "A review of time-series forecasting algorithms for industrial manufacturing systems," *Machines*, vol. 12, p. 380, 6 Jun. 2024, ISSN: 2075-1702. DOI: 10.3390/machines12060380.

[27] F. Rustam, A. A. Reshi, A. Mehmood, S. Ullah, B.-W. On, W. Aslam, and G. S. Choi, "Covid-19 future forecasting using supervised machine learning models," *IEEE Access*, vol. 8, pp. 101 489–101 499, 2020, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2020.2997311.

[28] W. Xu, Z. Li, C. Cheng, and T. Zheng, "Data mining for unemployment rate prediction using search engine query data," *Service Oriented Computing and Applications*, vol. 7, no. 1, pp. 33–42, 2013, Cited by: 32. DOI: 10.1007/s11761-012-0122-2.

[Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84875424551&doi=10.1007%2fs11761-012-0122-2&partnerID=40&md5=8f1c946a6811816367e32c1ab842ea1a.

[29] Instituto Nacional de Estatística, *População residente por local de residência, sexo e grupo etário*, Accessed: 31 May 2025, 2024. [Online]. Available: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&indOcorrCod=0008273&xlang=pt.

[30] J. Leal, A. Martins, and P. Marujo, *Regresso ao futuro: Uma base de dados das projeções macroeconómicas e orçamentais para a economia portuguesa no século xxi*, Conselho das Finanças Públicas, 2024.

[31] Fundação Francisco Manuel dos Santos (FFMS). "Pordata." Accessed: 4 August 2025. (), [Online]. Available: https://www.pordata.pt/home.

[32] Observador. "Em três anos de troika há menos 26 mil professores, menos disciplinas e mais exames." Acessed: 14 September 2025. (Jun. 2014), [Online]. Available: https://observador.pt/2014/06/13/em-tres-anos-de-troika-ha-menos-26-mil-professores-menos-disciplinas-e-mais-exames/.

[33] J. Fattah, L. Ezzine, Z. Aman, H. E. Moussami, and A. Lachhab, "Forecasting of demand using arima model," *International Journal of Engineering Business Management*, vol. 10, 2018, ISSN: 18479790. DOI: 10.1177/1847979018808673.

[34] E. A. Ilmani, F. F. Amatullah, K. A. Notodiputro, Y. Angraini, and L. N. A. Mualifah, "The performance of the arimax model on cooking oil price data in indonesia," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 19, pp. 819–828, 2 Apr. 2025, ISSN: 2615-3017. DOI: 10.30598/barekengvol19iss2pp819-828.

[35] T. Chen and C. Guestrin, "Xgboost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Aug. 2016, pp. 785–794, ISBN: 9781450342322. DOI: 10.1145/2939672.2939785.

[36] Governo da República Portuguesa. "Governo e parceiros sociais assinam novo acordo de concertação social." Acessed: 10 July 2025. (Oct. 2024), [Online]. Available: https://www.portugal.gov.pt/pt/gc24/comunicacao/noticia?i=governo-e-parceiros-sociais-assinam-novo-acordo-de-concertacao-social.

[37] S. Liu, L. Cheng, P. Chen, S. Xu, J. Gao, and J. Lu, "Prediction of industrial hazardous waste production based on different models," *IOP Conference Series: Earth and Environmental Science*, vol. 384, p. 012 026, 1 Nov. 2019, ISSN: 1755-1307. DOI: 10.1088/1755-1315/384/1/012026.

[38] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009, ISBN: 1441412697.

[39] W. McKinney, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, S. van der Walt and J. Millman, Eds., 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[40] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Rio, M. Wiebe, P. Peterson, P. Gerard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. DOI: 10.1038/s41586-020-2649-2. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2.

[41] C. C. Eric Gazoni, *Openpyxl - a python library to read/write excel 2010 xlsx/xlsm files*, 2024. [Online]. Available: https://openpyxl.readthedocs.io.

[42] K. Jordahl, J. V. den Bossche, M. Fleischmann, J. Wasserman, J. McBride, J. Gerard, J. Tratner, M. Perry, A. G. Badaracco, C. Farmer, G. A. Hjelle, A. D. Snow, M. Cochran, S. Gillies, L. Culbertson, M. Bartos, N. Eubank, maxalbert, A. Bilogur, S. Rey, C. Ren, D. Arribas-Bel, L. Wasser, L. J. Wolf, M. Journois, J. Wilson, A. Greenhall, C. Holdgraf, Filipe, and F. Leblanc, *Geopandas/geopandas: V0.8.1*, version v0.8.1, Jul. 2020. DOI: 10.5281/zenodo.3946761. [Online]. Available: https://doi.org/10.5281/zenodo.3946761.

[43] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[44] T. G. Smith *et al.*, *pmdarima: Arima estimators for Python*, 2017. [Online]. Available: http://www.alkaline-ml.com/pmdarima.

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[46] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. DOI: 10.1109/MCSE.2007.55.

[47] M. L. Waskom, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021. DOI: 10.21105/joss.03021. [Online]. Available: https://doi.org/10.21105/joss.03021.

[48] G. Van Rossum, *The Python Library Reference, release 3.8.2*. Python Software Foundation, 2020.

[This page is intentionally left blank.]

APPENDIX A

Best-performing models, per Category and Age Group

Table A.1. Best performing TS models, per Category and Age Group

Category	**Data Type**	<24	25–34	35–44	45–54	55–64	>65	**Total**
Education Ministry	Real Data	SARIMAX	SARIMAX	VAR	VAR	VAR	SARIMAX	ARIMAX
	Projected Data	SARIMAX	SARIMAX	ARIMAX	ARIMAX	SARIMAX	VAR	SARIMAX
Early Childhood + Primary/Secondary Teachers	Real Data	SARIMAX	VAR	ARIMAX	VAR	VAR	VAR	VAR
	Projected Data	ARIMAX	ARIMAX	ARIMAX	VAR	SARIMAX	SARIMAX	SARIMAX
Polytechnic HE Teachers	Real Data	ARIMAX	SARIMAX	VAR	VAR	ARIMAX	SARIMAX	VAR
	Projected Data	ARIMAX	ARIMAX	SARIMAX	SARIMAX	ARIMAX	VAR	ARIMAX
University Teachers	Real Data	ARIMAX	SARIMAX	VAR	VAR	VAR	ARIMAX	SARIMAX
	Projected Data	ARIMAX	ARIMAX	VAR	SARIMAX	ARIMAX	VAR	ARIMAX

Table A.2. Best-performing ML models, per Category and Age Group

Category	**Data Type**	<24	25–34	35–44	45–54	55–64	>65	**Total**
Education Ministry	Real Data	MLP	MLP	MLP	MLP	Grad. Boost.	MLP	Grad. Boost.
	Projected Data	MLP	MLP	MLP	MLP	Grad. Boost.	MLP	Grad. Boost.
Early Childhood + Primary/Secondary Teachers	Real Data	MLP	MLP	MLP	Grad. Boost.	Grad. Boost.	MLP	Grad. Boost.
	Projected Data	MLP	MLP	MLP	Grad. Boost.	Grad. Boost.	MLP	Random Forest Regressor
Polytechnic HE Teachers	Real Data	MLP	Grad. Boost.	Grad. Boost.	MLP	MLP	MLP	MLP
	Projected Data	MLP	MLP	Grad. Boost.	MLP	MLP	MLP	Grad. Boost.
University Teachers	Real Data	MLP	MLP	MLP	MLP	Grad. Boost.	MLP	MLP
	Projected Data	MLP	MLP	MLP	SVR	Grad. Boost.	MLP	MLP

Table A.3. Best-performing TS models with a fixed forecast period of 5 semesters, per Category and Age Group

Category	<24	25–34	35–44	45–54	55–64	>65	**Total**
Education Ministry	SARIMAX	ARIMAX	ARIMAX	ARIMAX	SARIMAX	ARIMAX	SARIMAX
Early Childhood + Primary/Secondary Teachers	ARIMAX	ARIMAX	ARIMAX	ARIMAX	SARIMAX	ARIMAX	SARIMAX
Polytechnic HE Teachers	ARIMAX	ARIMAX	VAR	SARIMAX	ARIMAX	VAR	VAR
University Teachers	ARIMAX	ARIMAX	VAR	SARIMAX	ARIMAX	VAR	ARIMAX

APPENDIX B

Detailed Future Forecast Model Performance and Predictions

Table B.1. Projected Work Positions (rounded to units), per Category and Age Group (2025–2027)

Category	Semester	<24	25–34	35–44	45–54	55–64	>65	Total	Sum of Age Groups
Education Ministry	2025	1,797	15,311	39,098	86,638	83,738	15,302	243,284	241,884
	2025.5	2,098	16,324	37,879	87,415	84,418	15,428	246,453	243,562
	2026	2,099	17,265	36,571	88,164	83,838	16,446	249,379	244,383
	2026.5	2,402	18,272	35,444	88,995	84,092	16,531	253,162	245,736
	2027	2,403	19,169	34,088	89,775	82,883	16,706	256,886	245,024
	2027.5	2,681	20,031	32,755	90,550	82,774	17,030	261,815	245,821
Early Childhood + Primary/Secondary Teachers	2025	625	4,868	16,559	50,549	49,866	8,959	130,942	131,426
	2025.5	722	5,177	15,183	50,877	50,370	9,367	130,229	131,696
	2026	783	5,514	14,304	51,257	50,172	9,999	130,290	132,029
	2026.5	906	5,831	12,990	51,599	50,516	10,331	129,195	132,173
	2027	1116	6,174	12,442	51,684	50,174	9,952	128,309	131,542
	2027.5	1192	6,523	11,317	52,061	50,412	10,475	126,873	131,980
Polytechnic HE Teachers	2025	127	1,095	2,754	4,526	3,133	362	12,457	11,997
	2025.5	125	1,175	2,918	4,753	3,221	351	12,765	12,543
	2026	131	1,157	2,733	4,662	3,319	365	12,557	12,367
	2026.5	130	1,219	2,884	4,920	3,416	333	12,210	12,902
	2027	128	1,217	2,690	4,781	3,500	328	10,933	12,644
	2027.5	132	1,267	2,833	5,022	3,597	306	10,817	13,157
University Teachers	2025	351	2,284	3,880	5,019	5,236	1,314	17,991	18,084
	2025.5	406	2,351	4,108	5,120	5,264	1,364	18,362	18,613
	2026	384	2,397	4,350	5,162	5,298	1,397	18,558	18,988
	2026.5	418	2,458	4,502	5,291	5,306	1,458	18,925	19,433
	2027	412	2,507	4,654	5,363	5,340	1,502	19,109	19,778
	2027.5	436	2,571	4,894	5,498	5,357	1,555	19,381	20,311

Table B.2. Absolute (rounded to 2 decimals) and Normalised Values (rounded to 3 decimals) in TS Approach with a fixed forecast period of 5 semesters, per Contract Type and NUTS III Region - Part 1

NUTS III	Data Type	Fixed Term	Permanent	Commission Service
Alentejo Central	Absolute	12.58	16.52	0.68
	Normalised	0.032	0.008	0.060
Alentejo Litoral	Absolute	23.19	21.57	1.11
	Normalised	0.071	0.021	0.109
Alto Alentejo	Absolute	39.99	30.46	0.61
	Normalised	0.109	0.023	0.045
Baixo Alentejo	Absolute	31.76	31.93	1.81
	Normalised	0.064	0.022	0.181
Algarve	Absolute	102.55	175.01	1.23
	Normalised	0.062	0.032	0.044
Beira Baixa	Absolute	8.82	12.03	0.33
	Normalised	0.053	0.011	0.048
Beiras e Serra da Estrela	Absolute	16.58	57.13	0.66
	Normalised	0.041	0.024	0.037
Região de Aveiro	Absolute	43.53	27.60	0.83
	Normalised	0.065	0.006	0.048
Região de Coimbra	Absolute	40.16	64.66	1.45
	Normalised	0.047	0.013	0.062
Região de Leiria	Absolute	54.84	69.50	1.39
	Normalised	0.111	0.022	0.098
Viseu Dão Lafões	Absolute	46.79	185.34	0.67
	Normalised	0.102	0.056	0.034
Grande Lisboa	Absolute	233.28	242.61	8.90
	Normalised	0.040	0.014	0.107
Alto Minho	Absolute	83.06	55.20	0.84
	Normalised	0.147	0.021	0.064

Table B.3. Absolute (rounded to 2 decimals) and Normalised Values (rounded to 3 decimals) in TS Approach with a fixed forecast period of 5 semesters, per Contract Type and NUTS III Region - Part 2

NUTS III	Data Type	Fixed Term	Permanent	Commission Service
Alto Tâmega e Barroso	Absolute	17.27	52.37	0.39
	Normalised	0.088	0.052	0.088
Área Metropolitana do Porto	Absolute	134.18	210.73	2.85
	Normalised	0.041	0.012	0.039
Ave	Absolute	79.18	37.24	1.83
	Normalised	0.104	0.008	0.075
Cávado	Absolute	65.87	129.20	0.53
	Normalised	0.075	0.027	0.027
Douro	Absolute	46.51	101.20	0.88
	Normalised	0.109	0.042	0.060
Tâmega e Sousa	Absolute	86.00	167.44	0.28
	Normalised	0.084	0.035	0.011
Terras de Trás-os-Montes	Absolute	21.76	113.58	0.91
	Normalised	0.083	0.089	0.106
Lezíria do Tejo	Absolute	31.22	30.30	0.63
	Normalised	0.044	0.011	0.047
Médio Tejo	Absolute	25.45	23.79	0.70
	Normalised	0.041	0.010	0.056
Oeste	Absolute	80.96	30.63	0.58
	Normalised	0.095	0.008	0.033
Península de Setúbal	Absolute	89.40	131.20	1.13
	Normalised	0.045	0.016	0.037
Região Autónoma da Madeira	Absolute	25.42	41.96	-
	Normalised	0.051	0.008	-
Região Autónoma dos Açores	Absolute	58.30	24.65	-
	Normalised	0.065	0.005	-

Table B.4. Additional Features relevance in TS Approach with a fixed forecast period of 5 semesters for Fixed-Term Contracts, per NUTS III Region

Region	Minimum Wage	GDP	Public Debt	GFCF	Number of Registered Students	Unemployment Rate
Alentejo Central	x	–	–	–	–	–
Alentejo Litoral	–	x	–	–	x	–
Alto Alentejo	x	–	x	x	x	–
Baixo Alentejo	x	–	x	–	–	–
Algarve	x	x	–	x	–	–
Beira Baixa	–	–	–	–	x	x
Beiras e Serra da Estrela	–	–	x	–	–	–
Região de Aveiro	–	–	–	–	–	x
Região de Coimbra	x	x	–	x	–	–
Região de Leiria	–	x	–	–	x	–
Viseu Dão Lafões	–	–	x	–	x	x
Grande Lisboa	x	–	x	–	–	x
Alto Minho	–	x	–	–	x	–
Alto Tâmega e Barroso	x	–	–	–	x	x
Área Metropolitana do Porto	–	x	–	–	x	-
Ave	–	x	–	–	x	–
Cávado	x	x	x	–	x	–
Douro	–	x	–	–	x	–
Tâmega e Sousa	–	–	–	x	x	–
Terras de Trás-os-Montes	–	x	x	x	–	–
Lezíria do Tejo	–	x	–	–	x	–
Médio Tejo	–	x	x	x	–	–
Oeste	–	–	–	x	x	x
Península de Setúbal	x	–	x	x	–	x
Região Autónoma da Madeira	x	x	–	–	–	–
Região Autónoma dos Açores	–	–	–	–	x	–

Table B.5. Additional Features relevance in TS Approach with a fixed forecast period of 5 semesters for Permanent Contracts, per NUTS III Region

Region	Minimum Wage	GDP	Public Debt	GFCF	Number of Registered Students	Unemployment Rate
Alentejo Central	x	x	–	x	–	x
Alentejo Litoral	x	–	x	x	–	–
Alto Alentejo	x	–	–	x	–	x
Baixo Alentejo	x	–	x	–	–	x
Algarve	x	–	–	x	x	–
Beira Baixa	–	x	–	–	–	x
Beiras e Serra da Estrela	x	–	–	–	x	–
Região de Aveiro	–	x	x	x	–	x
Região de Coimbra	x	–	–	–	–	–
Região de Leiria	–	–	x	x	–	x
Viseu Dão Lafões	x	–	–	–	–	–
Grande Lisboa	x	–	–	x	–	–
Alto Minho	–	x	–	x	–	x
Alto Tâmega e Barroso	x	x	–	x	–	–
Área Metropolitana do Porto	–	–	x	x	x	–
Ave	–	–	x	x	–	x
Cávado	–	x		–	–	–
Douro	x	x	x	–	x	x
Tâmega e Sousa	x	–	x	x	x	–
Terras de Trás-os-Montes	x	x	–	x	–	x
Lezíria do Tejo	x	x	x	–	–	x
Médio Tejo	–	x	–	–	x	x
Oeste	–	–	x	–	–	–
Península de Setúbal	–	–	x	–	x	–
Região Autónoma da Madeira	–	–	–	–	–	x
Região Autónoma dos Açores	x	–	x	x	–	x

Table B.6. Additional Features relevance in TS Approach with a fixed forecast period of 5 semesters for Commission Service or Political Office/Mandate Contracts, per NUTS III Region

Region	Minimum Wage	GDP	Public Debt	GFCF	Number of Registered Students	Unemployment Rate
Alentejo Central	–	–	**x**	–	–	**x**
Alentejo Litoral	**x**	**x**	**x**	–	**x**	–
Alto Alentejo	**x**	–	–	**x**	–	**x**
Baixo Alentejo	**x**	–	**x**	–	**x**	**x**
Algarve	–	–	–	–	–	**x**
Beira Baixa	**x**	–	–	–	–	**x**
Beiras e Serra da Estrela	**x**	–	–	–	–	–
Região de Aveiro	**x**	–	**x**	**x**	–	**x**
Região de Coimbra	–	–	–	–	**x**	–
Região de Leiria	–	–	–	**x**	**x**	–
Viseu Dão Lafões	**x**	–	**x**	–	**x**	**x**
Grande Lisboa	**x**	–	**x**	–	**x**	–
Alto Minho	**x**	–	**x**	–	**x**	–
Alto Tâmega e Barroso	**x**	**x**	**x**	**x**	**x**	–
Área Metropolitana do Porto	**x**	–	**x**	–	–	**x**
Ave	**x**	**x**	**x**	–	–	–
Cávado	**x**	**x**	**x**	–	–	**x**
Douro	**x**	**x**	**x**	–	**x**	–
Tâmega e Sousa	**x**	**x**	–	**x**	–	**x**
Terras de Trás-os-Montes	–	–	–	**x**	–	**x**
Lezíria do Tejo	**x**	**x**	**x**	–	–	–
Médio Tejo	**x**	–	**x**	–	**x**	**x**
Oeste	**x**	**x**	**x**	-	–	**x**
Península de Setúbal	**x**	**x**	**x**	–	–	**x**

Table B.7. Projected Work Positions (rounded to units), per NUTS III Region and Contract Type (2025–2027) - Part 1

NUTS III	Semester	Fixed Term	Permanent	Commission Service	Sum
Alentejo Central	2025	351	2,125	11	2,487
	2025.5	240	2,169	11	2,420
	2026	266	2,168	10	2,444
	2026.5	142	2,193	10	2,345
	2027	155	2,203	10	2,368
	2027.5	19	2,238	10	2,267
Alentejo Litoral	2025	253	1,158	10	1,421
	2025.5	244	1,269	11	1,524
	2026	326	1,366	10	1,702
	2026.5	344	1,486	11	1,841
	2027	418	1,659	12	2,089
	2027.5	362	1,764	12	2,138
Alto Alentejo	2025	272	1,454	12	1,738
	2025.5	169	1,594	13	1,776
	2026	185	1,647	14	1,846
	2026.5	82	1,816	12	1,910
	2027	106	1,903	11	2,020
	2027.5	1	2,097	7	2,105
Baixo Alentejo	2025	361	1,595	12	1,968
	2025.5	165	1,712	11	1,888
	2026	130	1,756	11	1,897
	2026.5	-114	1,888	11	1,899
	2027	-190	1,951	10	1,961
	2027.5	-488	2,101	8	2,109
Algarve	2025	1,267	6,161	29	7,457
	2025.5	1,000	6,794	30	7,824
	2026	877	7,069	28	7,974
	2026.5	554	7,828	27	8,409
	2027	147	8,354	27	8,528
	2027.5	-144	9,191	28	9,219
Beira Baixa	2025	169	1,094	7	1,270
	2025.5	113	1,130	8	1,251
	2026	98	1,128	5	1,231
	2026.5	18	1,185	4	1,207
	2027	-22	1,192	5	1,197
	2027.5	-129	1,251	5	1,256
Beiras e Serra da Estrela	2025	383	2,462	18	2,863
	2025.5	322	2,623	18	2,963
	2026	343	2,722	18	3,083
	2026.5	281	2,876	18	3,175
	2027	302	3,015	18	3,335
	2027.5	241	3,247	18	3,506
Região de Aveiro	2025	653	4,460	10	5,123
	2025.5	571	4,521	13	5,105
	2026	664	4,511	11	5,186
	2026.5	614	4,586	11	5,211
	2027	731	4,692	10	5,433
	2027.5	702	4,738	10	5,450

Table B.8. Projected Work Positions (rounded to units), per NUTS III Region and Contract Type (2025–2027) - Part 2

NUTS III	Semester	Fixed Term	Permanent	Commission Service	Sum
Região de Coimbra	2025	857	4,994	30	5,881
	2025.5	767	5,271	30	6,068
	2026	866	5,293	27	6,186
	2026.5	791	5,630	27	6,448
	2027	831	5,700	28	6,559
	2027.5	802	6,058	25	6,885
Região de Leiria	2025	466	3,107	18	3,591
	2025.5	556	3,050	20	3,626
	2026	578	2,986	13	3,577
	2026.5	562	2,927	9	3,498
	2027	662	2,938	7	3,607
	2027.5	542	2,904	20	3,466
Viseu Dão Lafões	2025	457	3,637	21	4,115
	2025.5	313	4,008	19	4,340
	2026	454	4,311	20	4,785
	2026.5	327	4,826	21	5,174
	2027	519	5,265	22	5,806
	2027.5	377	5,891	21	6,289
Grande Lisboa	2025	3,969	19,569	112	23,650
	2025.5	1,901	21,797	123	23,821
	2026	1,404	23,123	134	24,661
	2026.5	-836	25,949	146	26,095
	2027	-1,602	28,084	157	28,241
	2027.5	-3,954	31,425	167	31,592
Alto Minho	2025	294	2,635	10	2,939
	2025.5	579	2,704	9	3,292
	2026	566	2,640	9	3,215
	2026.5	670	2,723	8	3,401
	2027	927	2,769	7	3,703
	2027.5	649	2,822	2	3,473
Alto Tâmega e Barroso	2025	179	1,097	3	1,279
	2025.5	122	1,206	3	1,331
	2026	140	1,276	3	1,419
	2026.5	97	1,409	3	1,509
	2027	127	1,539	2	1,668
	2027.5	60	1,692	2	1,754
Área Metropolitana do Porto	2025	3,645	16,538	66	20,249
	2025.5	3,091	16,038	65	19,194
	2026	4,211	14,592	64	18,867
	2026.5	3,514	13,509	63	17,086
	2027	4,419	11,696	61	16,176
	2027.5	3,604	9,974	59	13,637
Ave	2025	614	4,507	25	5,146
	2025.5	716	4,662	24	5,402
	2026	855	4,549	24	5,428
	2026.5	971	4,683	24	5,678
	2027	1,086	4,569	24	5,679
	2027.5	948	4,694	24	5,666

Table B.9. Projected Work Positions (rounded to units), per NUTS III Region and Contract Type (2025–2027) - Part 3

NUTS III	Semester	Fixed Term	Permanent	Commission Service	Sum
Cávado	2025	957	4,903	21	5,881
	2025.5	750	5,207	20	5,977
	2026	925	5,096	20	6,041
	2026.5	752	5,383	19	6,154
	2027	936	5,279	19	6,234
	2027.5	768	5,584	18	6,370
Douro	2025	563	2,491	13	3,067
	2025.5	357	2,503	12	2,872
	2026	665	2,509	13	3,187
	2026.5	372	2,520	12	2,904
	2027	732	2,534	13	3,279
	2027.5	268	2,535	11	2,814
Tâmega e Sousa	2025	1,050	4,672	26	5,748
	2025.5	1,028	4,731	24	5,783
	2026	1,053	4,576	24	5,653
	2026.5	1,008	4,613	23	5,644
	2027	820	4,471	22	5,313
	2027.5	829	4,455	21	5,305
Terras de Trás-os-Montes	2025	234	1,372	8	1,614
	2025.5	208	1,580	9	1,797
	2026	214	1,749	7	1,970
	2026.5	168	1,942	10	2,120
	2027	154	2,171	9	2,334
	2027.5	117	2,493	10	2,620
Lezíria do Tejo	2025	713	3,065	11	3,789
	2025.5	650	3,306	11	3,967
	2026	933	3,407	10	4,350
	2026.5	816	3,685	10	4,511
	2027	1,017	3,823	9	4,849
	2027.5	712	4,121	8	4,841
Médio Tejo	2025	685	2,329	7	3,021
	2025.5	677	2,425	4	3,106
	2026	746	2,288	0	3,034
	2026.5	780	2,489	-4	3,269
	2027	822	2,236	-9	3,058
	2027.5	845	2,605	-14	3,450
Oeste	2025	880	3,964	17	4,861
	2025.5	869	4,270	16	5,155
	2026	891	4,370	16	5,277
	2026.5	867	4,749	15	5,631
	2027	742	4,921	15	5,678
	2027.5	754	5,368	14	6,136
Península de Setúbal	2025	1,727	8,774	34	10,535
	2025.5	1,178	9,324	36	10,538
	2026	1,198	9,223	37	10,458
	2026.5	586	9,806	38	10,430
	2027	426	9,711	39	10,176
	2027.5	-155	10,275	38	10,313

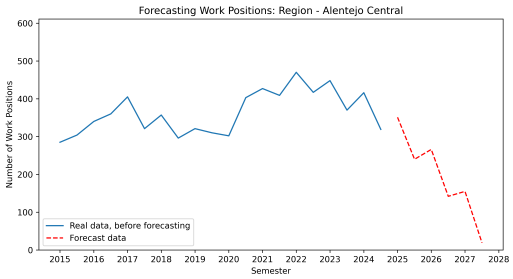
Table B.10. Projected Work Positions (rounded to units), per NUTS III Region and Contract Type (2025–2027) - Part 4

NUTS III	Semester	Fixed Term	Permanent	Commission Service	Sum
Região Autónoma da Madeira	2025	396	4,944	-	5,340
	2025.5	415	4,858	-	5,273
	2026	524	4,766	-	5,290
	2026.5	508	4,676	-	5,184
	2027	614	4,586	-	5,200
	2027.5	664	4,497	-	5,161
Região Autónoma dos Açores	2025	756	4,806	-	5,562
	2025.5	657	4,873	-	5,530
	2026	647	4,952	-	5,599
	2026.5	505	5,042	-	5,547
	2027	466	5,134	-	5,600
	2027.5	238	5,207	-	5,445

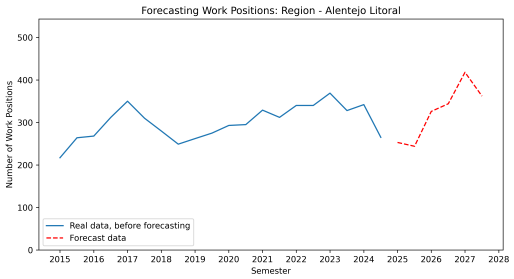
APPENDIX C

Future Forecast Visualisations of Workforce Positions

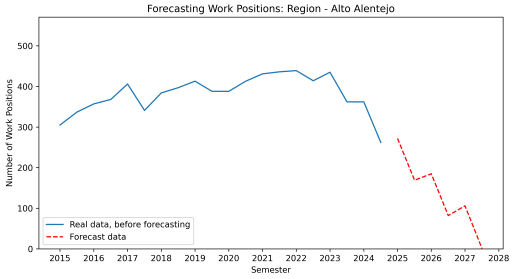
C.1. Future Work Positions Predictions, for Fixed-Term Contracts



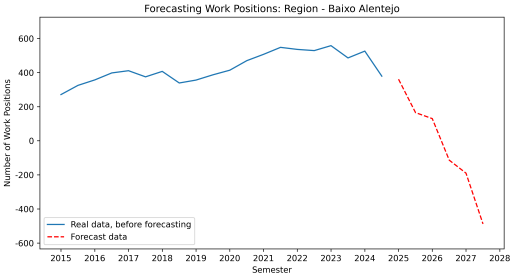
(a) Region: *Alentejo Central*



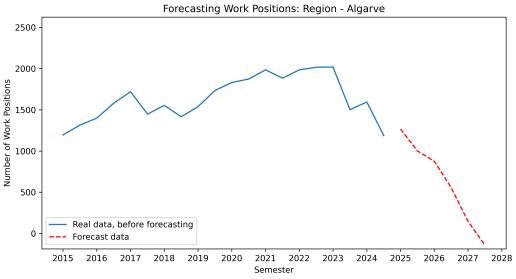
(b) Region: *Alentejo Litoral*



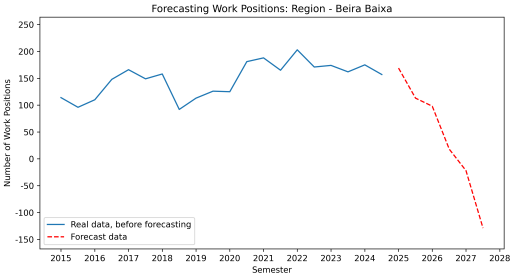
(c) Region: *Alto Alentejo*



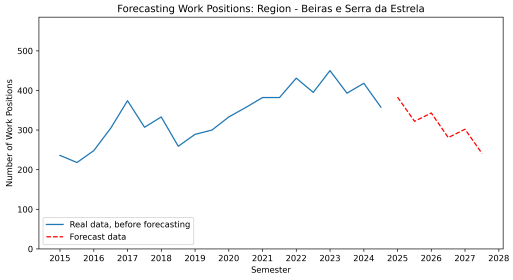
(d) Region: *Baixo Alentejo*



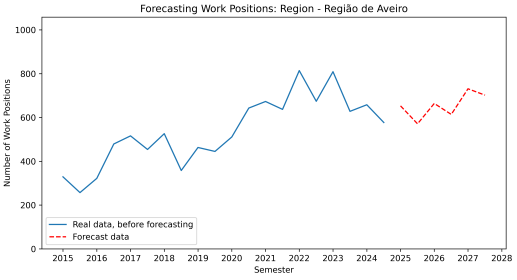
(e) Region: *Algarve*



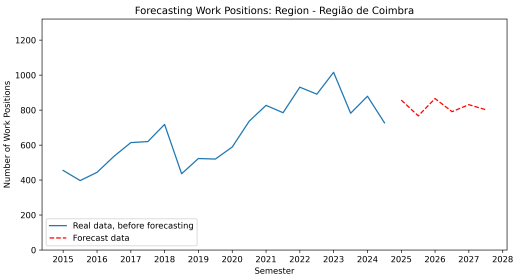
(f) Region: *Beira Baixa*



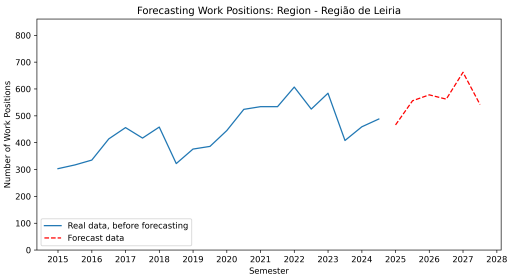
(g) Region: *Beiras e Serra da Estrela*



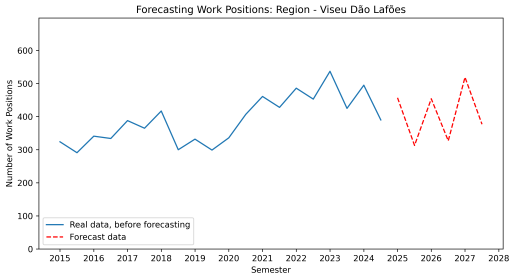
(h) Region: *Região de Aveiro*



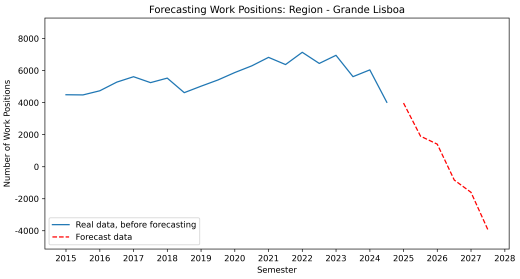
(i) Region: *Região de Coimbra*



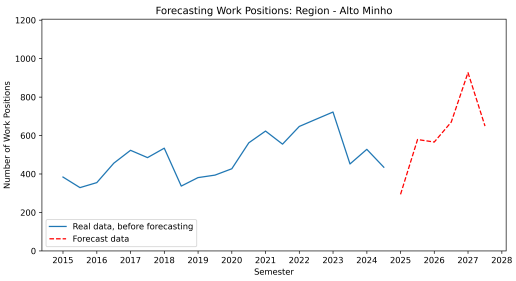
(j) Region: *Região de Leiria*



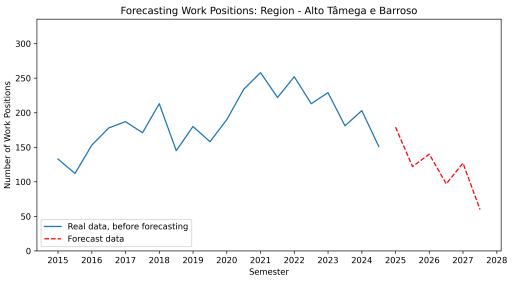
(k) Region: *Viseu Dão Lafões*



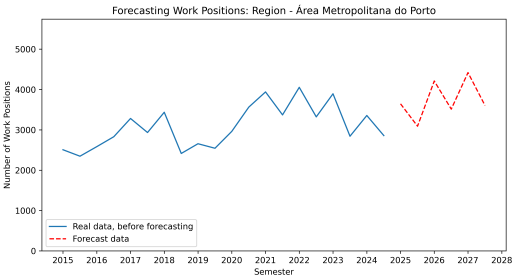
(l) Region: *Grande Lisboa*



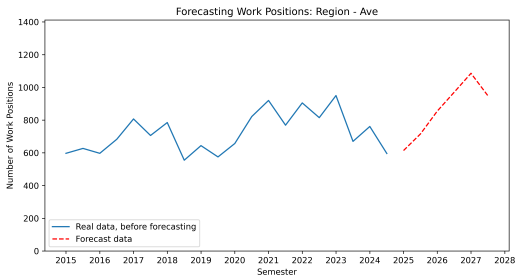
(m) Region: *Alto Minho*



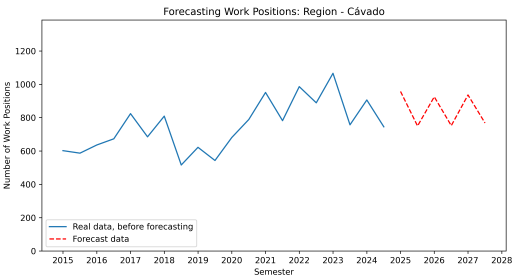
(n) Region: *Alto Tâmega e Barroso*



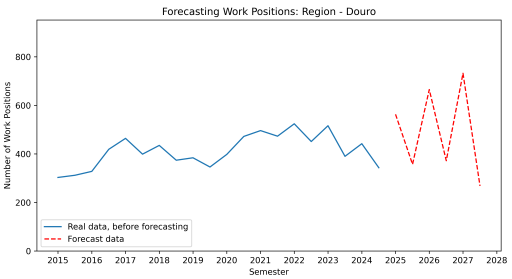
(o) Region: *Área Metropolitana do Porto*



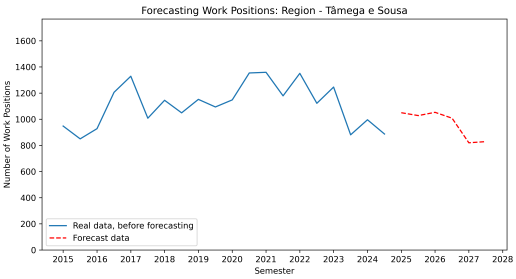
(p) Region: *Ave*



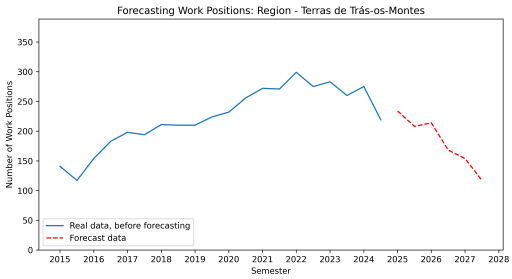
(q) Region: *Cávado*



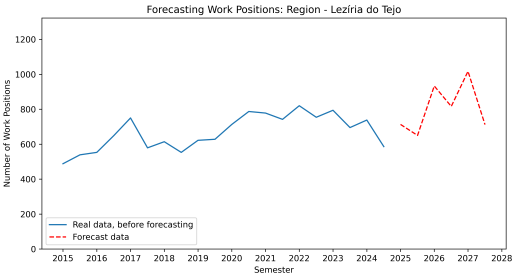
(r) Region: *Douro*



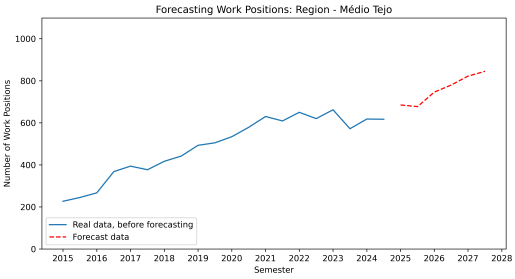
(s) Region: *Tâmega e Sousa*



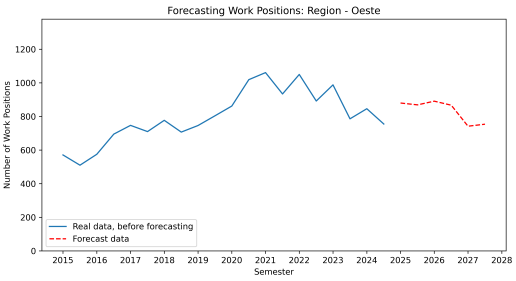
(t) Region: *Terras de Trás-os-Montes*



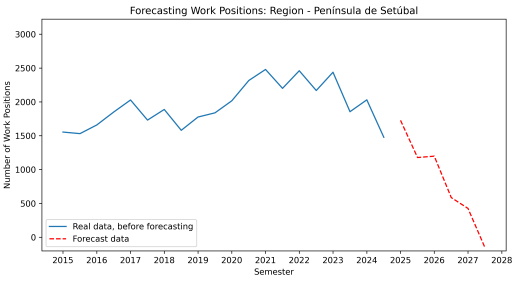
(u) Region: *Lezíria do Tejo*



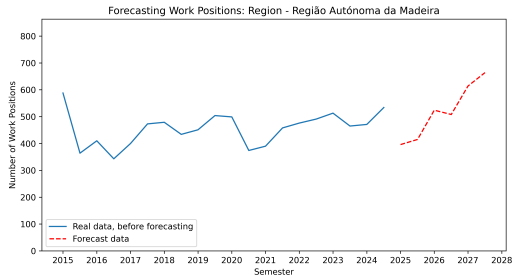
(v) Region: *Médio Tejo*



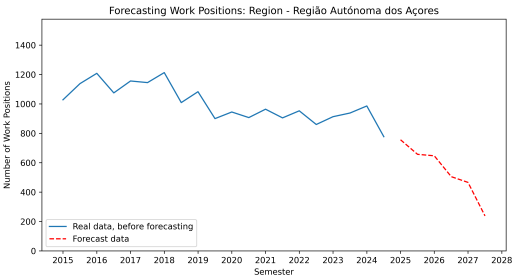
(w) Region: *Oeste*



(x) Region: *Península de Setúbal*

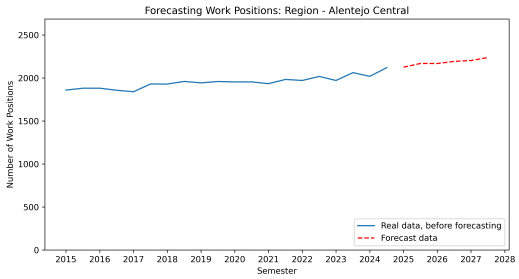


(y) Region: *Região Autónoma da Madeira*

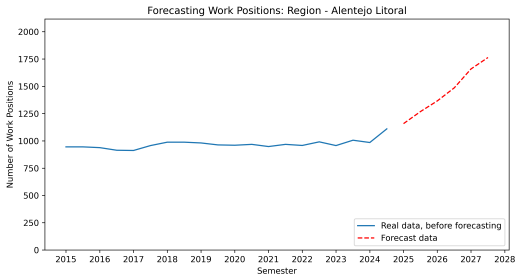


(z) Region: *Região Autónoma dos Açores*

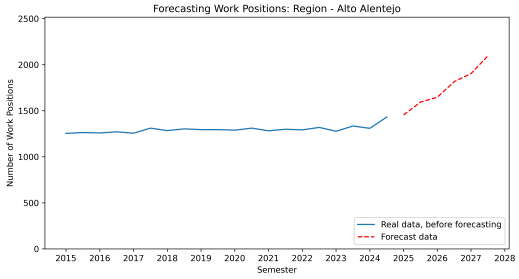
C.2. Future Work Positions Predictions, for Permanent Contracts



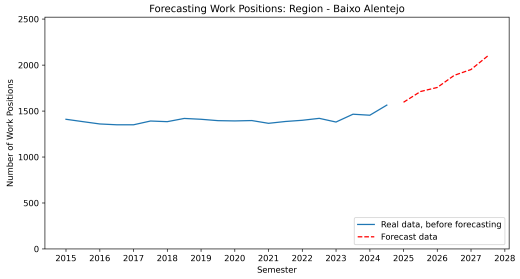
(a) Region: *Alentejo Central*



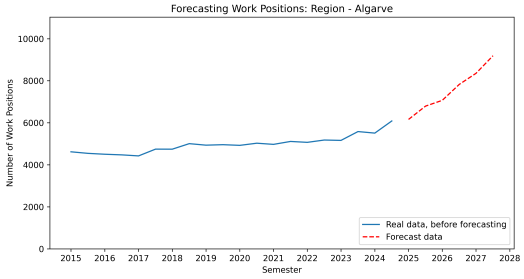
(b) Region: *Alentejo Litoral*



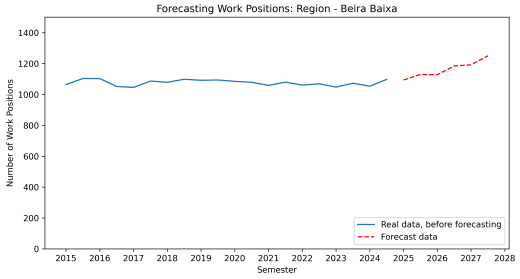
(c) Region: *Alto Alentejo*



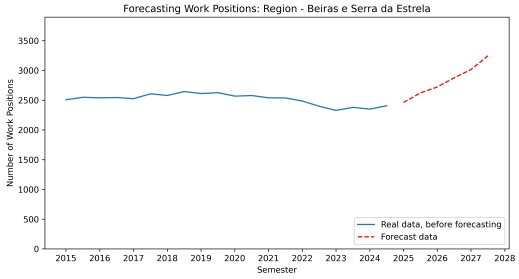
(d) Region: *Baixo Alentejo*



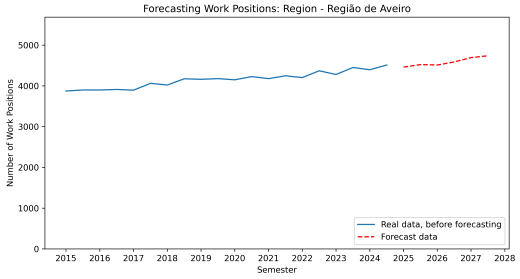
(e) Region: *Algarve*



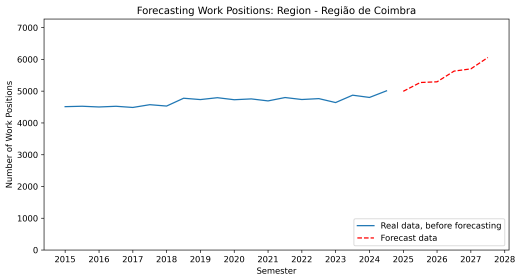
(f) Region: *Beira Baixa*



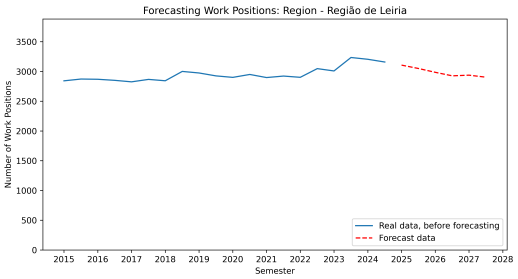
(g) Region: *Beiras e Serra da Estrela*



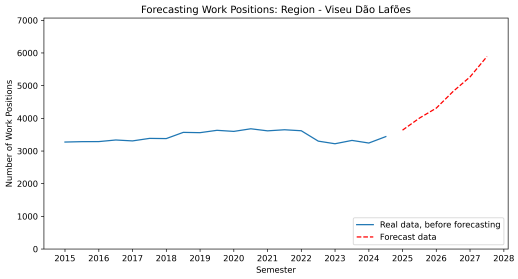
(h) Region: *Região de Aveiro*



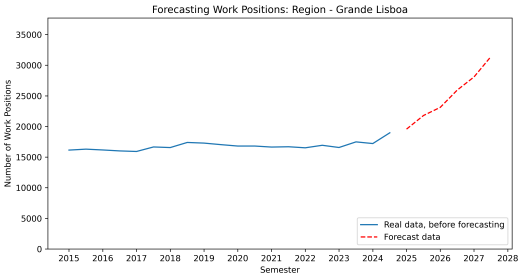
(i) Region: *Região de Coimbra*



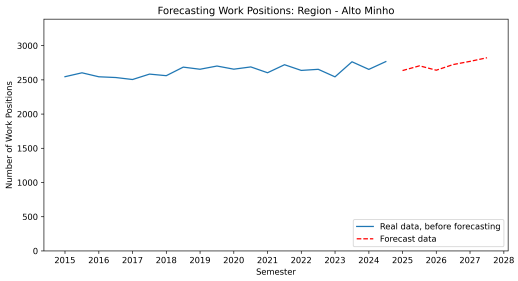
(j) Region: *Região de Leiria*



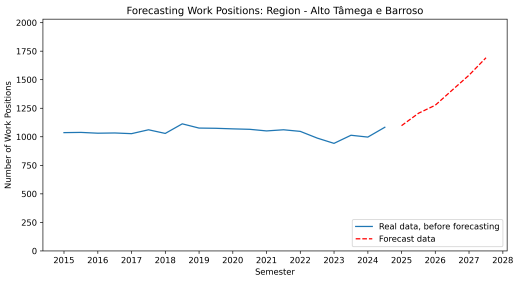
(k) Region: *Viseu Dão Lafões*



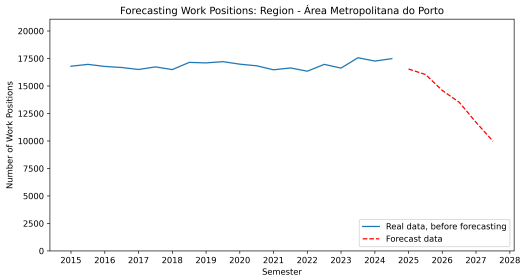
(l) Region: *Grande Lisboa*



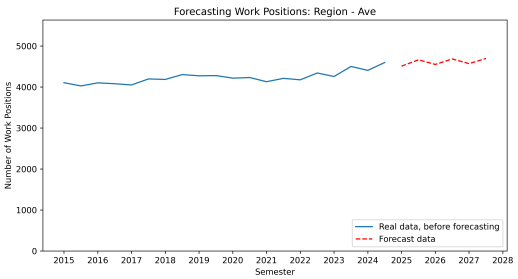
(m) Region: *Alto Minho*



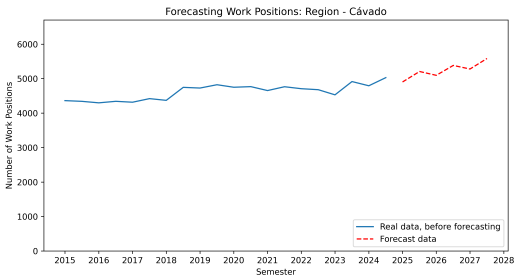
(n) Region: *Alto Tâmega e Barroso*



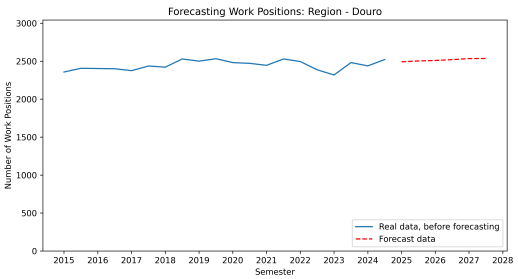
(o) Region: *Área Metropolitana do Porto*



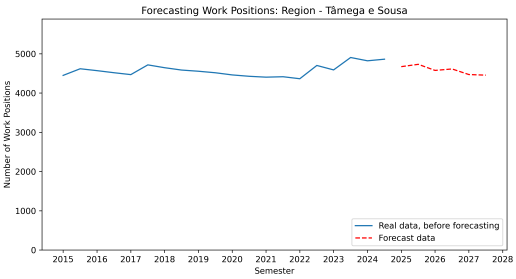
(p) Region: *Ave*



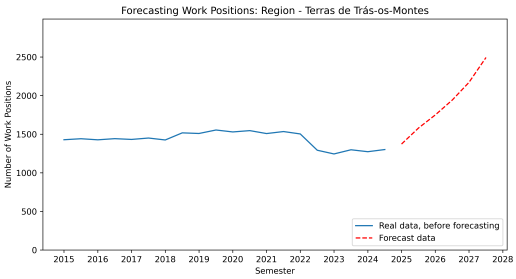
(q) Region: *Cávado*



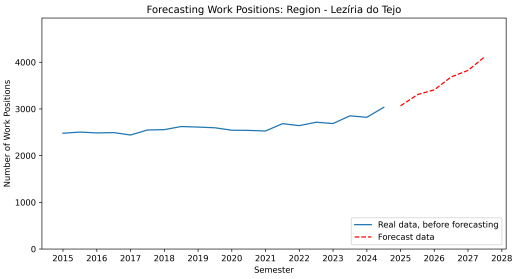
(r) Region: *Douro*



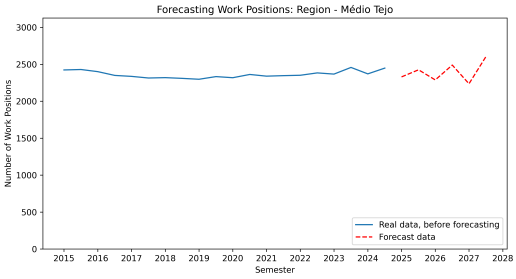
(s) Region: *Tâmega e Sousa*



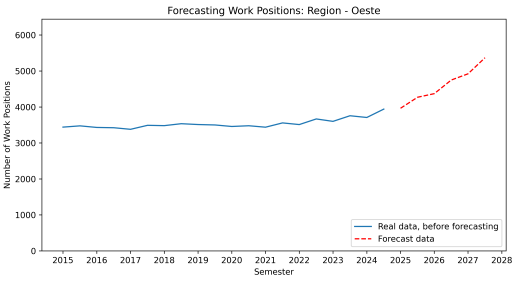
(t) Region: *Terras de Trás-os-Montes*



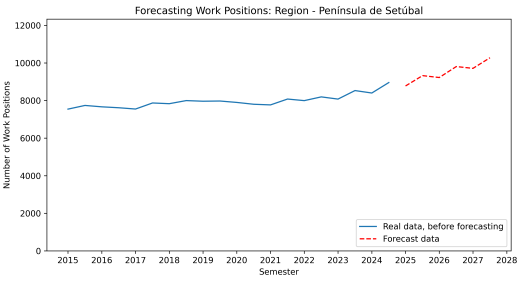
(u) Region: *Lezíria do Tejo*



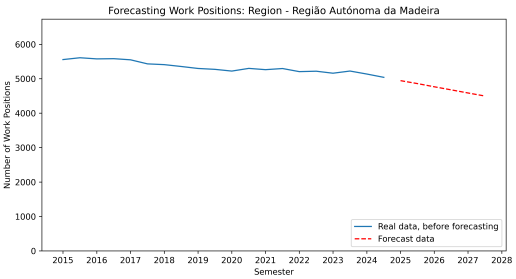
(v) Region: *Médio Tejo*



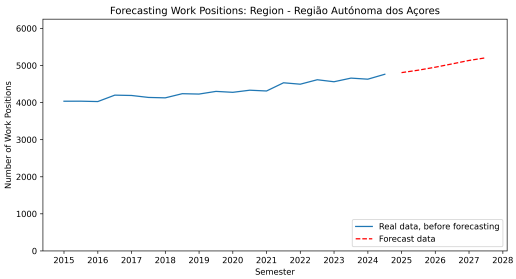
(w) Region: *Oeste*



(x) Region: *Península de Setúbal*

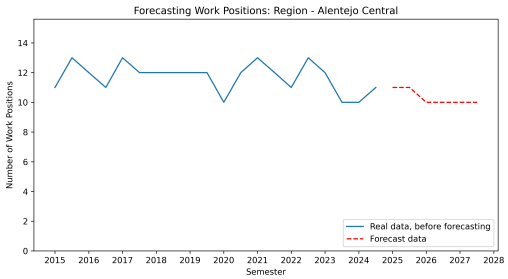


(y) Region: *Região Autónoma da Madeira*

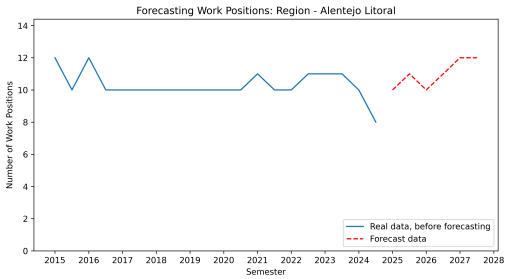


(z) Region: *Região Autónoma dos Açores*

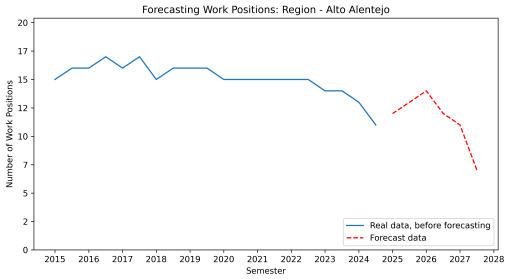
C.3. Future Work Positions Predictions, for Commission Service or Political Office/Mandate Contracts



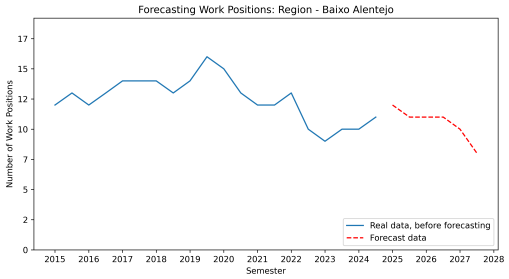
(a) Region: *Alentejo Central*



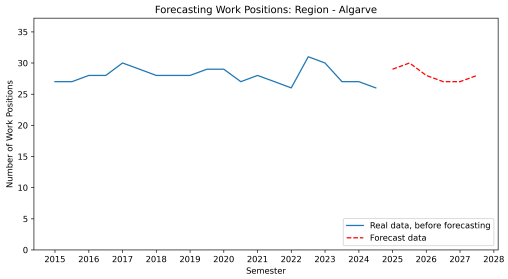
(b) Region: *Alentejo Litoral*



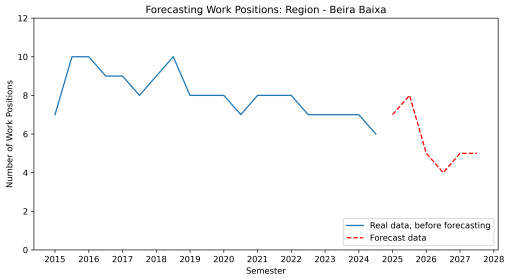
(c) Region: *Alto Alentejo*



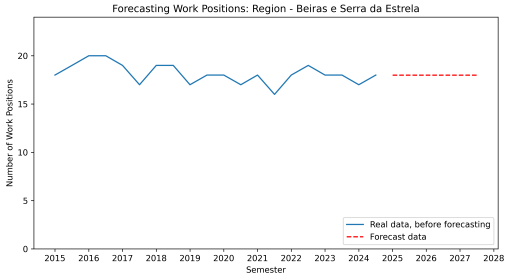
(d) Region: *Baixo Alentejo*



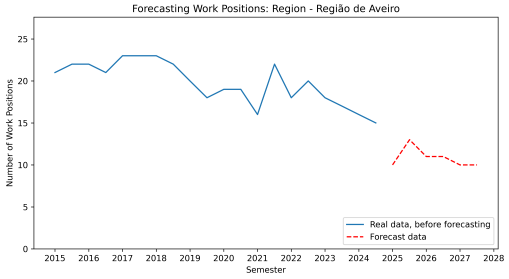
(e) Region: *Algarve*



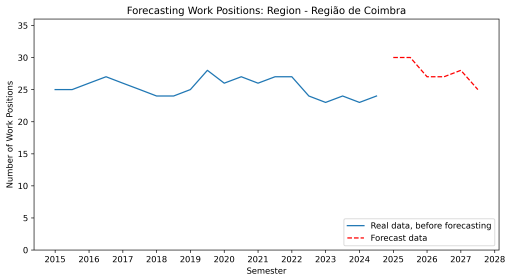
(f) Region: *Beira Baixa*



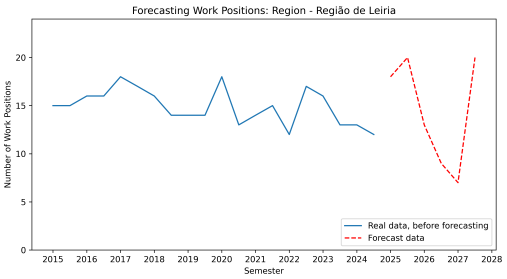
(g) Region: *Beiras e Serra da Estrela*



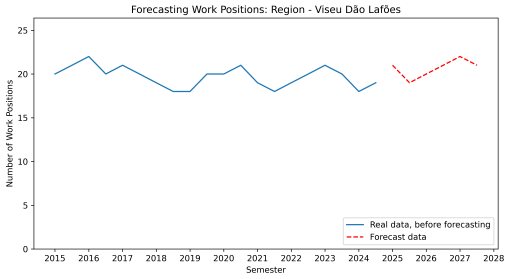
(h) Region: *Região de Aveiro*



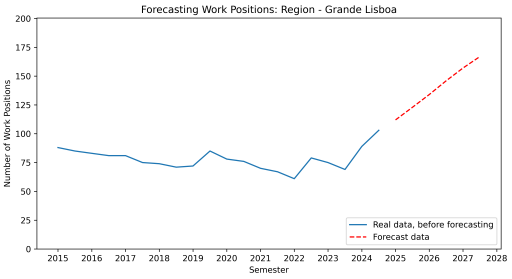
(i) Region: *Região de Coimbra*



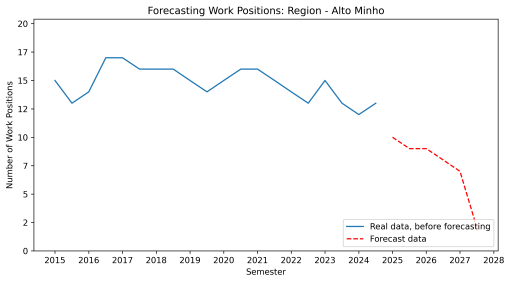
(j) Region: *Região de Leiria*



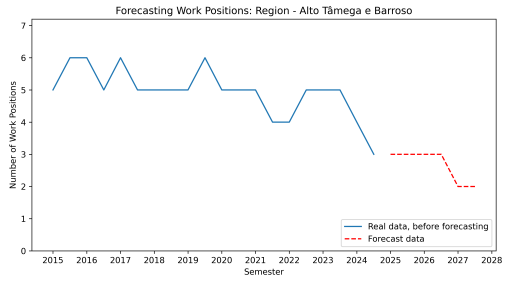
(k) Region: *Viseu Dão Lafões*



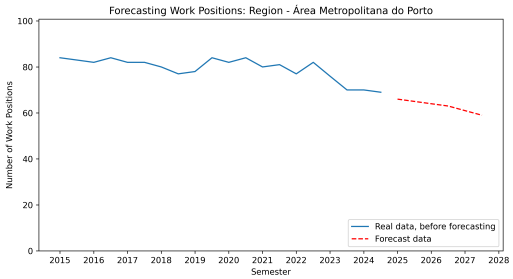
(l) Region: *Grande Lisboa*



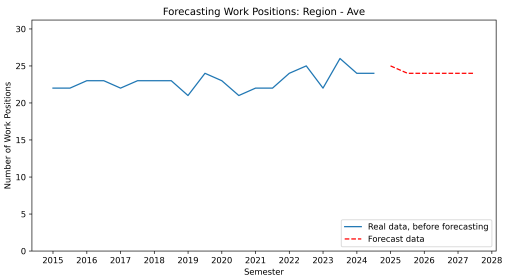
(m) Region: *Alto Minho*



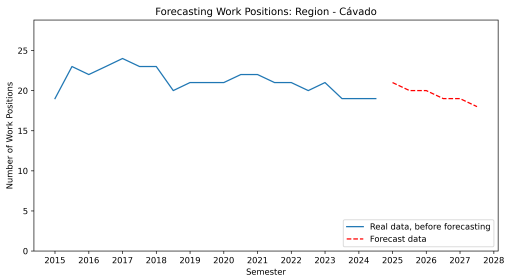
(n) Region: *Alto Tâmega e Barroso*



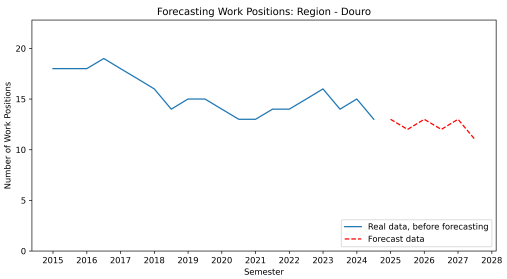
(o) Region: *Área Metropolitana do Porto*



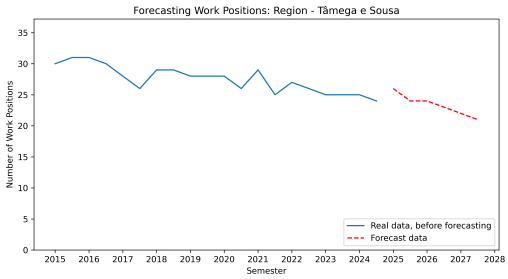
(p) Region: *Ave*



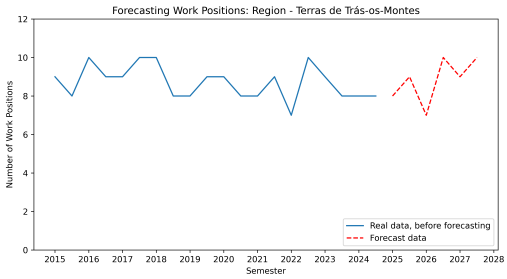
(q) Region: *Cávado*



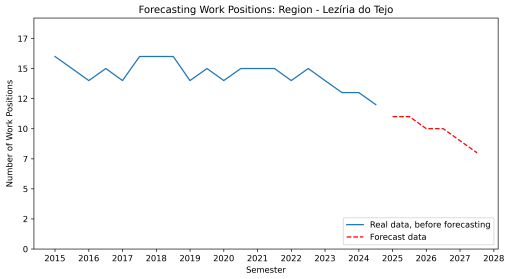
(r) Region: *Douro*



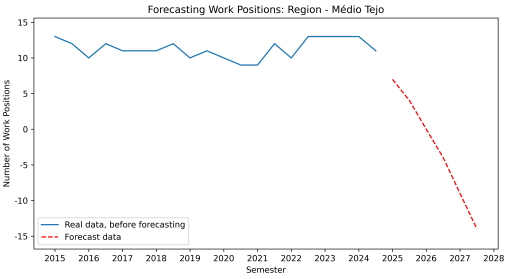
(s) Region: *Tâmega e Sousa*



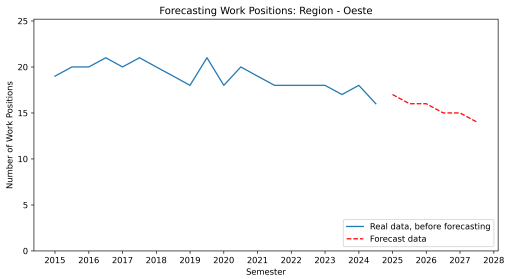
(t) Region: *Terras de Trás-os-Montes*



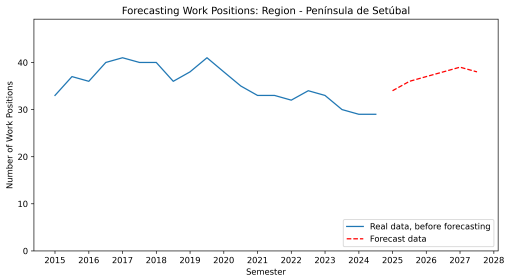
(u) Region: *Lezíria do Tejo*



(v) Region: *Médio Tejo*



(w) Region: *Oeste*



(x) Region: *Península de Setúbal*