How to Explain the NEETs distribution in the European Union: Comparing the Performance of Regression Trees and Linear Regression Models

Introduction

The Not in Employment, Education or Training (NEET) indicates the share of young people aged 15 to 29 who do not engage in formal employment, education, or training programmes (OECD, 2024). The NEET average across OECD countries is 12.5% and 11% in the EU (Eurostat, 2024). Variations across states and regions are quite high, with best performing countries like the Netherlands with a share of NEETs below 5% and worst performing countries like Romania and Italy above 15% (Eurostat, 2024). The percentage of NEET can be a powerful indicator of the vulnerabilities of young citizens when transitioning from education to work or to assess the performance of labour markets.

In this report, we use a regional panel dataset constructed from publicly available Eurostat indicators, to investigate what variables best predict the share of NEET in different EU regions. We compare the performance and the results of a regression tree and linear regressions model. Our empirical analysis addresses the three research questions: (1) identifying socio-economic and educational factors associated with regional NEET rates across Europe, (2) exploring potential non-linear thresholds, and (3) comparing the performance of regression trees with linear regression models.

We find that regional NEET rates are strongly associated with labour-market strength, educational attainment, and gender inequality, but that these relationships are not purely linear. Regression trees outperform linear models in out-of-sample prediction and reveal clear threshold effects, particularly in the gender employment gap and early school leaving, that are not easily detected in linear specifications. These results suggest that combining parametric and machine-learning approaches yields a richer understanding of regional youth disengagement and provides policy-relevant insights into the conditions under which improvements in education and labour markets translate into meaningful reductions in NEET rates.

Literature Review

A large and consistent body of empirical research identifies labour market conditions and educational attainment as the primary determinants of NEET outcomes. In a systematic review and meta-analysis of 43 studies, Rahmani et al. (2024) document strong associations between NEET status and socio-economic, educational, and demographic factors. Prior unemployment experience substantially increases the likelihood of being NEET, while higher levels of education act as a protective factor. These findings align with earlier descriptive and analytical work emphasising the central role of youth unemployment, inactivity, and early school leaving, particularly during economic downturns (Logez, 2013).

Education-related variables are especially influential. Early leavers from education face significantly higher risks of long-term disengagement, while tertiary attainment and

participation in lifelong learning are associated with lower NEET rates by improving employability and adaptability (Logez, 2012; Miluka & Meurs, 2024). Still, the effectiveness of education systems depends on their alignment with labour market needs. Countries and regions with well-developed vocational education and training, apprenticeships, and work-based learning schemes tend to experience smoother school-to-work transitions and lower NEET rates, although these effects weaken during periods of economic crisis (Miluka & Meurs, 2024). Rahmani et al. (2024) also highlight pronounced gender disparities and heterogeneous patterns across demographic cohorts, suggesting that the impact of education and employment conditions on NEET outcomes is not uniform.

At the regional level, spatial heterogeneity and persistence in NEET rates have been documented in recent studies. Maynou et al. (2022) analyse NEET dynamics across 274 European regions and identify distinct convergence clubs characterised by different long-run trajectories. While overall regional disparities have declined over time, substantial heterogeneity remains. Across all clusters, unemployment rates and early school leaving consistently emerge as the main drivers of NEET rates, but their relative importance varies across regions. This evidence suggests that similar changes in labour market or educational conditions can yield different NEET outcomes depending on regional context, pointing to regime-specific effects.

Despite this rich empirical literature, most existing studies rely on linear regression frameworks that impose constant marginal effects across regions and levels of development. Such approaches may obscure non-linearities and threshold effects implied by both micro-level evidence (Rahmani et al., 2024) and regional convergence analyses (Maynou et al., 2022). For instance, improvements in employment rates or income levels may only translate into sharp reductions in NEET rates once critical levels are reached, while reductions in early school leaving may exhibit diminishing returns.

For this reason, the present study contributes to the literature by combining regional panel data with both linear and non-linear modelling approaches to better understand youth disengagement across European regions. Hence, we compare two models in addressing the following research questions:

- RQ1. What socioeconomic, educational, and employment factors best explain differences in youth disengagement (NEET rates) across European regions?
- RQ2. Are there specific thresholds in income, education, or employment levels beyond which NEET rates decrease sharply across European regions?
- RQ3. Can regression tree models capture non-linearities in the determinants of NEET rates more effectively than linear regression models?

Methodology

Data and Sample Construction

This study relies on a regional panel dataset constructed from publicly available Eurostat indicators at the NUTS-2 level, covering European regions over the period 2010-2023. The unit of analysis is the region-year. The choice of the NUTS-2 level reflects a trade-off between geographical granularity and data availability, as this level captures meaningful regional heterogeneity while ensuring sufficient coverage across countries and years. The dependent variable is the share of young people neither in employment nor in education or training (NEET) aged 15-29, expressed as a percentage of the corresponding population. This indicator captures youth disengagement from both the labour market and the education system and is widely used in policy and academic analyses of social exclusion.

We focus on a core set of explanatory variables capturing economic conditions, labour market structure, educational attainment, and gender inequality, selected on the basis of prior literature and data availability. These include regional GDP per capita (in purchasing power standards), the employment rate of the population aged 20-64, the long-term unemployment rate, the share of early leavers from education and training (aged 18–24), tertiary educational attainment among young adults (aged 25-34), adult participation in lifelong learning (aged 25-64), and the gender employment gap. To account for potential structural changes associated with the COVID-19 pandemic, we also include a post-2020 dummy variable equal to one for years from 2020 onwards. All variables are harmonised to a common regional and temporal structure. GDP per capita is transformed using the natural logarithm to reduce skewness and allow for diminishing marginal effects. The final estimation sample is restricted to region-year observations with complete data on the core variables, yielding a balanced dataset suitable for both linear and non-linear modelling approaches.

Exploratory Analysis and Multicollinearity Diagnostics

Prior to model estimation, we conduct an exploratory data analysis to assess the distribution of NEET rates across regions and over time, and to examine bivariate relationships between the dependent variable and key covariates. Correlation matrices reveal strong associations between NEET rates and several labour market and education variables, suggesting the potential presence of multicollinearity. To formally assess this issue, we compute Variance Inflation Factors (VIFs) for alternative model specifications. The results indicate moderate correlation among some regressors, particularly between employment-related and education-related variables, but VIF values remain below conventional thresholds, suggesting that multicollinearity does not critically compromise estimation. Based on these diagnostics, we proceed with two complementary model specifications: one emphasising labour market conditions and another focusing on educational determinants.

Linear Regression Framework

As a baseline, we estimate multiple linear regression models using Ordinary Least Squares (OLS). These models provide a global, parametric benchmark for assessing the average association between regional characteristics and NEET rates. Given the presence of heteroskedasticity in the residuals, all linear models are estimated with heteroskedasticity-robust (HC1) standard errors. While the linear specification offers ease of interpretation and statistical inference, it implicitly assumes that the effect of each explanatory variable is constant across regions and levels of development. This assumption may be overly restrictive in the presence of regional heterogeneity and threshold effects, motivating the use of non-linear methods.

Regression Trees and Non-linear Modelling

To capture potential non-linearities and interaction effects, we complement the linear analysis with regression tree models. Regression trees recursively partition the covariate space into homogeneous regions by selecting threshold values that minimize within-node variance of the outcome variable. This approach allows for the identification of discrete regimes and critical thresholds beyond which NEET rates change sharply. To balance predictive performance and interpretability, we estimate a parsimonious, interpretable tree constrained by a limited depth and a minimum number of observations per leaf. The latter facilitates substantive interpretation of the splits and aligns with the policy-oriented objectives of the analysis.

Train–Test Split and Overfitting Control

To mitigate concerns related to overfitting, particularly relevant for flexible, non-parametric models, we adopt a train-test split strategy. The dataset is randomly divided into a training sample (80%) and a test sample (20%). All model estimation and threshold identification are conducted exclusively on the training data, while predictive performance is evaluated out-of-sample on the test data. Model performance is assessed using multiple metrics, including the Root Mean Squared Error (RMSE), the Mean Absolute Error (MAE), and the coefficient of determination (R^2). Comparing in-sample and out-of-sample performance allows us to evaluate the extent of overfitting and to assess whether the patterns identified by the regression tree generalise beyond the estimation sample.

Model Comparison and Interpretation

The comparison between linear regression and regression tree models serves two complementary purposes. First, it allows us to assess whether accounting for non-linearities improves predictive accuracy. Second, it provides insights into the structure of regional heterogeneity in youth disengagement by identifying salient thresholds in key determinants such as gender employment gaps, income levels, and early school leaving rates.

Results

NEET rates display substantial cross-regional variation over 2010–2023. In the complete-case estimation sample, the mean NEET rate is 16.3% (s.d. 8.6), with values ranging from 3.1% to 53.8%, indicating large disparities in youth disengagement across European regions. Covariates also show meaningful dispersion: employment rates average 68.1% (s.d. 10.0), early school leaving averages 13.7% (s.d. 9.9) with a long upper tail, and the gender employment gap is sizable (mean 14.8 p.p., s.d. 11.3).

A descriptive NUTS-2 choropleth for 2017 reveals clear geographic clustering. Higher NEET rates concentrate in parts of Southern and South-East Europe, while lower rates are more common in many Northern and Central regions. For cartographic clarity, French outermost regions (FRY*) are excluded, as they lie outside continental Europe and would otherwise distort the map's scale.

Bivariate correlations point to a coherent set of strong relationships. NEET is highly negatively associated with the employment rate ($r = -0.93$) and log GDP per capita ($r = -0.74$), and positively associated with early school leaving ($r = 0.70$) and the gender employment gap ($r = 0.72$). These patterns suggest that regional labour-market strength, overall economic development, and human-capital indicators are tightly linked to youth disengagement. Long-term unemployment is more weakly correlated with NEET ($r = 0.27$), hinting that its role may be more conditional on other regional characteristics.

The OLS models provide a parametric benchmark for the direction and magnitude of associations. In the labour-market specification (Model A), NEET rates are significantly lower in regions with higher employment and higher income, and higher where long-term unemployment and the gender employment gap are larger. The model fits the data closely ($R^2 = 0.884$), with the employment rate emerging as the strongest correlate in magnitude. The post-2020 indicator is positive and statistically significant, consistent with a shift towards higher disengagement after 2020 conditional on observed covariates.

In the education-oriented specification (Model B), the associations remain economically meaningful but the model explains less variance ($R^2 = 0.751$). NEET is higher in regions with greater early school leaving and larger gender employment gaps, and lower where adult learning participation is higher and GDP per capita is greater. Tertiary attainment enters with a small positive coefficient; given the aggregate regional nature of the data, this likely reflects compositional or structural differences correlated with tertiary concentration rather than a direct protective mechanism at the individual level. Across both models, robust (HC1) inference confirms that the core correlates are precisely estimated.

Out-of-sample evaluation confirms that the linear benchmark generalizes well: OLS performance is stable across training and test sets ($R^2\_train = 0.753$, $R^2\_test = 0.742$; RMSE = 4.31 vs 4.22).

Regression trees improve predictive accuracy, particularly out-of-sample. A cross-validated tuned tree achieves substantially higher fit in the training data (R^2_train = 0.909) while maintaining a clear gain on the test set (R^2_test = 0.806; test RMSE 3.66). The train–test gap indicates some overfitting, but the improvement on unseen data suggests that the tree captures relevant non-linear structure not well approximated by a global linear specification.

To extract policy-relevant thresholds, we estimate a constrained "interpretable" tree (depth 3; minimum leaf size 100). The resulting partitions are driven primarily by three predictors: the gender employment gap, log GDP per capita, and early school leaving. Consistent with this, feature-importance comparisons show that these variables account for nearly all splitting power in the interpretable tree, while tertiary attainment and adult learning play a limited role in defining regimes once the main partitions are in place.

The split points imply discrete thresholds around which NEET levels differ markedly: the gender employment gap is repeatedly split near 13.6–18.8 p.p., early school leaving near 13.8–19.6%, and log GDP per capita near 9.7–10.2. Visualizations corroborate sharp differences in NEET distributions across these thresholds: regions with relatively low gender gaps and low early school leaving systematically exhibit lower NEET rates. Importantly, these regime contrasts remain visible when stratifying observations by train vs test membership, supporting the external validity of the main threshold patterns.

Taken together, the results indicate that youth disengagement across European regions is tightly associated with labour-market performance and inequality-related factors, and that these relationships are not purely linear. Regression trees both enhance predictive performance and provide an interpretable segmentation of regional contexts, highlighting threshold-like regimes, especially in the gender employment gap and early school leaving, within which NEET rates differ sharply.

Discussion: Linking Results to the Research Questions

Addressing RQ1, our results provide strong and consistent evidence that labour-market conditions, educational outcomes, and gender inequalities are the primary factors explaining cross-regional variation in youth disengagement across Europe.

Across both linear specifications, NEET rates are strongly negatively associated with overall employment performance and regional economic development, as proxied by the employment rate of the working-age population and GDP per capita. Regions with stronger labour markets and higher income levels exhibit substantially lower NEET rates, highlighting the central role of demand-side conditions and macroeconomic context in shaping youth outcomes.

Educational factors also emerge as key correlates. In the education-focused model, early school leaving shows a large and positive association with NEET rates, while adult participation in lifelong learning is associated with lower youth disengagement. These

findings suggest that both initial educational trajectories and opportunities for skill accumulation later in life contribute to regional resilience against youth inactivity.

Finally, the gender employment gap consistently appears as one of the strongest predictors of NEET rates across all model specifications. Regions with larger disparities between male and female employment exhibit significantly higher levels of youth disengagement, underscoring the importance of gendered labour-market structures and social norms in shaping aggregate youth outcomes.

Regarding RQ2, the regression tree analysis provides clear evidence of non-linearities and threshold effects in the determinants of NEET rates. Unlike linear models, the tree identifies discrete regimes in which NEET outcomes change sharply once key variables cross specific levels. Three thresholds stand out as particularly salient. First, the gender employment gap repeatedly defines the initial splits of the interpretable tree, with critical values around 13-19 percentage points. Regions below these thresholds exhibit substantially lower NEET rates than those above them, suggesting that reductions in gender inequality beyond a certain point yield disproportionately large gains in youth engagement.

Second, early school leaving acts as a decisive educational threshold. Regions with early school leaving rates below roughly 14-20% experience markedly lower NEET rates, whereas regions above these levels fall into high-disengagement regimes. This finding highlights the existence of tipping points in educational attainment, beyond which youth disengagement becomes structurally entrenched. Third, income levels, measured by log GDP per capita, further segment regions within these regimes. Thresholds around log GDP values of 9.7-10.2 indicate that the protective effect of income is not linear: once a minimum level of economic development is reached, additional income is associated with smaller marginal reductions in NEET rates.

Finally, considering our RQ3, the comparison between linear regression and regression trees shows that non-linear models capture the structure of NEET determinants more effectively, both in predictive and substantive terms.

From a predictive standpoint, regression trees outperform OLS in out-of-sample evaluation. While linear models generalise well, the tree achieves a higher test-set R^2 and lower prediction errors, indicating that allowing for interactions and non-linear effects improves explanatory power. Although some degree of overfitting is observed in-sample, the tree's superior test performance suggests that the identified patterns generalise beyond the estimation sample.

From an interpretative perspective, regression trees offer insights that are not accessible through linear specifications. By explicitly identifying thresholds and interaction structures, the tree reveals how combinations of gender inequality, educational outcomes, and income jointly define high- and low-NEET regimes. These regime-based insights are particularly

valuable from a policy perspective, as they point to critical margins where interventions may have the greatest impact.

Thus, the evidence strongly supports RQ3: regression trees provide a more nuanced and informative representation of the determinants of youth disengagement than linear models, complementing rather than replacing traditional parametric approaches.

Together, the results show that (i) youth disengagement is shaped by a multidimensional set of socioeconomic and educational factors; (ii) these relationships are characterised by sharp thresholds rather than smooth gradients; and (iii) regression trees are especially well suited to uncovering this non-linear structure. This integrated framework directly addresses the three research questions and strengthens the case for combining linear and machine-learning approaches in regional labour-market analysis.

Conclusion and Future Research

This paper analyses regional variation in youth disengagement across Europe by combining linear regression models with non-linear regression trees. Using a NUTS-2 panel dataset for 2010–2023, we show that NEET rates are strongly associated with labour-market performance, educational outcomes, and gender inequality, consistent with existing empirical evidence. However, our results also reveal that these relationships are not purely linear.

The regression tree analysis identifies clear threshold effects, particularly in the gender employment gap and early school leaving rates, beyond which NEET outcomes improve sharply. These non-linear patterns are not fully captured by standard linear models, which, while informative, impose constant marginal effects across regions. From both a predictive and interpretative perspective, regression trees therefore provide valuable complementary insights by uncovering discrete regimes of youth disengagement.

These findings have direct policy implications. They suggest that marginal improvements in employment or education may have limited effects in regions that remain above critical thresholds, whereas targeted interventions addressing structural inequalities, especially gender disparities and early school leaving, can generate disproportionately large reductions in NEET rates once key tipping points are reached.

Future research could extend this analysis by adopting causal identification strategies, incorporating institutional and policy variables, or exploring dynamic transitions between regional NEET regimes over time. Such extensions would further enhance our understanding of the mechanisms driving youth disengagement and inform more effective, region-specific policy responses.

References

Eurostat. (2024). *Statistics on young people neither in employment nor in education or training - Statistics Explained*. Ec.europa.eu. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Statistics_on_young_people_neither_in_employment_nor_in_education_or_training

Logez, K. (2013, May 29). *What makes a NEET?* OECD Education and Skills Today. https://oecdedutoday.com/what-makes-a-neet/

Maynou, L., Ordóñez, J., & Silva, J. I. (2022). Convergence and determinants of young people not in employment, education or training: An European regional analysis. *Economic Modelling*, *110*, 105808. https://doi.org/10.1016/j.econmod.2022.105808

Meurs, M., & Miluka, J. (2024). Gender Differences in the Determinants of Young NEETs: Evidence from Albania. *World Bank Group*.

OECD. (2024). *Youth not in employment, education or training (NEET)*. OECD. https://www.oecd.org/en/data/indicators/youth-not-in-employment-education-or-training-neet.html

Rahmani, H., Groot, W., & Rahmani, A. M. (2024). Unravelling the NEET phenomenon: A systematic literature review and meta-analysis of risk factors for youth not in education, employment, or training. *International Journal of Adolescence and Youth*, *29*(1). https://doi.org/10.1080/02673843.2024.2331576