HULT
INTERNATIONAL
BUSINESS SCHOOL

A3: Individual Assignment

Introduction to Machine Learning & AI

Angel Lanto

Master’s in Business Analytics

Table of Contents

Table of Figures

I. Introduction

a. Problem Definition

Skills mismatch refers to the inconsistency between the qualifications and skills possessed by individuals and those required by their jobs (Foxton, 2016). A common method of identifying this mismatch is by comparing a worker's educational attainment with the typical qualifications associated with their occupation. Overeducation and undereducation are two forms of mismatch where workers possess either higher or lower qualifications than their roles demand. This misalignment complicates labour market dynamics and challenges the effective deployment of human capital.

b. Significance

Understanding skills mismatch is vital due to its broader economic implications. A mismatch may indicate inefficient allocation of labour, where the potential of highly educated individuals is underutilised, or less qualified individuals may struggle in roles beyond their capacity. This inefficiency can suppress overall productivity, affect job satisfaction, and hinder wage growth. From a policy perspective, identifying skills mismatches helps inform education systems, training programmes, and employment services to better align the workforce with market demand. As highlighted by international bodies like the ILO and the UK Office for National Statistics (ONS), monitoring mismatch trends can support long-term economic planning and individual career outcomes.

c. Data Selection

The dataset presents the percentage distribution of educational mismatch in the UK labour market, covering the period from April–June 2002 to October–December 2015. It categorises workers as matched, overeducated, or undereducated, based on the alignment between their highest qualification and the typical educational requirement of their occupation. The data serves as a proxy measure of skills mismatch across multiple demographic and employment variables.

The dataset is disaggregated by the following variables:

- Gender (male, female)
- Age groups
- Employment type (full-time, part-time)
- Employment status (employee, self-employed)
- Country of birth (UK-born vs non-UK-born)

Each category includes percentage values for the proportion of workers classified as:

- Matched – those whose education level aligns with their occupation

- Overeducated – those with higher qualifications than typically required
- Undereducated – those with lower qualifications than typically required

This structured breakdown allows for temporal and comparative analysis of mismatch trends across socio-economic groups, providing insight into how efficiently skills have been allocated in the UK labour market over time. The data is sourced from the UK Office for National Statistics (ONS) and is based on Labour Force Survey (LFS) responses.

d. Relation of Problem to Machine Learning

This problem of skills mismatch aligns closely with several core topics in machine learning. Classification techniques can be applied to predict whether a worker is likely to be overeducated, undereducated, or well-matched, based on features such as age, gender, employment status, and country of birth. For example, a logistic regression model or a random forest classifier could be trained to identify key predictors of mismatch. Additionally, unsupervised learning methods such as clustering (e.g., k-means) could help uncover hidden patterns or groupings within the workforce that experience similar mismatch trends, which may not be evident through traditional analysis. If the dataset is expanded to include time-series elements, neural networks like recurrent neural networks (RNNs) could potentially model temporal trends in mismatch probabilities. Applying these techniques enables deeper insight into the structural causes of mismatch and supports data-driven recommendations for policy and education alignment.

II. Feature Engineering and Exploratory Data Analysis

To ensure the dataset was ready for meaningful modeling and visualization, the project began with meticulous feature engineering and exploratory data analysis (EDA). This process first involved cleaning and standardizing multiple Excel sheets containing skills mismatch data segmented by demographic and employment factors (e.g., age, gender, country of birth). Column names were formatted consistently, missing and duplicate records were removed, and non-numeric symbols were cleaned and converted for analysis. The cleaning step was necessary to ensure that downstream processes like modeling or visualization would operate on clean and reliable data.

During feature engineering, new columns were created to better reflect educational alignment in the labor market. These included matched_ratio (the proportion of workers whose qualifications matched job requirements) and education_gap (the difference between overeducated and undereducated individuals). These engineered features were essential for highlighting where imbalances were most pronounced and made the data more suitable for use in classification models and trend analyses.

Exploratory data analysis was conducted by grouping the data by segments (e.g., gender, employment type), allowing a close examination of numeric distributions (mean, min, max) and variability across groups. This analysis helped uncover patterns of mismatch that may be specific to certain population segments, guiding more targeted policy or organizational responses.

Note: Kindly hover over the plots in the Jupyter notebook to interactively explore values and compare segments across time. Additionally, this notebook was developed with ChatGPT (2025a & 2025b) as a guide, helping streamline the analytical process. It provided step-by-step support in designing code blocks, debugging logic, choosing appropriate modeling techniques, and shaping visualizations, essentially acting as a collaborative assistant to ensure clarity and depth in analysis.

III. Machine Learning Algorithms

To address the issue of skills mismatch, two machine learning algorithms which are Logistic Regression and K-Means Clustering, were evaluated for their suitability in uncovering patterns and classifying segments based on mismatch types. Both approaches offer distinct yet complementary strengths: one for prediction and the other for unsupervised segmentation.

Logistic Regression is a powerful linear classification model particularly well-suited for problems where the relationship between variables and outcomes is approximately linear. In the context of this project, engineered features such as matched_ratio and education_gap served as strong predictors of mismatch types (matched, overeducated, or undereducated). Logistic regression's ability to produce interpretable coefficients makes it valuable for understanding which factors contribute most to the likelihood of mismatch. Moreover, the model outputs probabilities, enabling stakeholders to assess confidence levels in predictions and prioritize interventions accordingly. Its simplicity, robustness, and suitability for structured tabular data made it an ideal candidate for the classification task.

On the other hand, K-Means Clustering provided an unsupervised approach to grouping individuals into segments based on similarity in mismatch characteristics, without relying on predefined labels. This method was useful in discovering hidden patterns across dimensions such as gender, age, employment type, and country of birth. Standardization was applied to ensure all features contributed equally to the distance calculations. The optimal number of clusters was determined using the Elbow Method, which helped balance granularity and interpretability. K-Means revealed which groups shared similar mismatch dynamics, offering insights that complemented the supervised model's predictions.

After comparing the two, Logistic Regression was selected as the final model for classification purposes. This decision was based on several factors: the presence of linearly related engineered features, the need for interpretability in policy contexts, and the model's resistance to overfitting when combined with techniques like L2 regularization and stratified k-fold cross-validation. Logistic regression not only predicted mismatch categories effectively but also helped quantify the relative importance of each feature, making it a practical choice for evidence-based decision-making.

Meanwhile, K-Means Clustering added exploratory value by segmenting the workforce into natural groups that may require tailored policy or training programs. Though not used as the final predictive model, K-Means provided crucial context and helped validate assumptions made during the classification stage.

In conclusion, while both models contributed uniquely to understanding skills mismatch, Logistic Regression was chosen as the primary algorithm due to its predictive strength, transparency, and alignment with the analytical goals. K-Means Clustering served as a supportive exploratory tool that enriched the interpretation of results and identified potential target segments for intervention strategies. Together, they formed a comprehensive approach to tackling education-to-employment disparities.

IV. Performance Improvement Strategies

To enhance model performance in analyzing skills mismatch, both logistic regression and K-Means clustering were optimized through a combination of feature engineering, data wrangling, regularization, hyperparameter tuning, and validation techniques. These enhancements ensured more accurate predictions, improved segmentation insights, and better generalization across various demographic and employment groups.

Feature engineering introduced two crucial variables: matched_ratio and education_gap, derived from the raw dataset. These features captured meaningful aspects of mismatch, such as the proportion of workers in correctly matched roles and the imbalance between overeducation and undereducation. These engineered metrics became central to both classification and clustering tasks, increasing model effectiveness by distilling complex patterns into quantifiable insights.

Data wrangling steps such as handling missing values, removing duplicates, and converting symbolic placeholders (e.g., ":") into valid numeric values ensured the model trained on clean, consistent data. The merging of multiple demographic sheets into a unified dataset with a segments column enabled better contextual interpretation across gender, age, work type, and country of origin.

For logistic regression, enhancements included L2 regularization to reduce overfitting and stratified k-fold cross-validation to maintain consistent class distribution across training and validation splits. Additionally, grid search was employed to optimize the regularization strength (C) and solver methods, ensuring the model was fine-tuned to the data's structure. These methods collectively improved the model's generalizability, stability, and predictive accuracy.

In parallel, K-Means clustering was applied as an unsupervised learning technique to uncover latent patterns among worker segments. To improve clustering performance, standardization of numerical features was performed using a StandardScaler, ensuring equal contribution from all variables. The Elbow Method was used to determine the optimal number of clusters by plotting the within-cluster sum of squares (WCSS) and identifying the point of diminishing returns. This helped prevent both under- and over-clustering. K-Means provided valuable groupings based on similarity in mismatch characteristics, offering a complementary perspective to supervised models.

The rationale for these enhancements lies in their ability to address the dataset's complexity and varied segment types. Regularization and cross-validation improved the robustness and interpretability of the logistic model, making it ideal for predicting mismatch categories. On the other hand, K-Means added value by segmenting the workforce into natural clusters without needing predefined labels—revealing, for example, which age or employment types shared similar mismatch dynamics.

In summary, the combination of logistic regression (for classification) and K-Means clustering (for segmentation), supported by robust preprocessing, feature engineering, and tuning techniques, resulted in a powerful dual-approach. This enhanced both the predictive power and interpretability of the skills mismatch analysis, enabling more informed and targeted policy or business decisions.

V. Ensemble Model

To enhance predictive accuracy and model robustness in analyzing skills mismatch, an ensemble learning approach was adopted, integrating Random Forest Classifier and XGBoost. These models were selected due to their superior performance in classification tasks involving structured tabular data and their ability to handle complex, non-linear relationships. The Random Forest model, built on the principle of bagging, aggregated predictions from multiple decision trees trained on different subsets of the dataset. This reduced variance and prevented overfitting, which was particularly beneficial given the segment-based nature of the data. It also offered feature importance metrics, confirming that engineered variables such as matched_ratio and education_gap were the most influential predictors.

In parallel, XGBoost, a gradient boosting method, was implemented for its capability to iteratively minimize classification errors and capture subtle patterns in imbalanced data. The model was optimized through hyperparameter tuning, including adjustments to learning rate, max depth, and number of estimators. Techniques such as early stopping and cross-validation were employed to further reduce overfitting and enhance generalization.

To combine the strengths of both models, a Voting Classifier was applied using soft voting to average the predicted probabilities. This stacked ensemble harnessed the stability and simplicity of Random Forest with the precision and adaptability of XGBoost. Together, they produced superior classification results, particularly in distinguishing minority mismatch categories like overeducated and undereducated, which often posed challenges in simpler models.

The ensemble approach significantly improved the model’s F1-score and overall accuracy, outperforming baseline classifiers such as logistic regression. It also provided deeper interpretability through feature rankings and probability distributions, which supported more nuanced insights into demographic or employment segments most at risk for mismatch.

In summary, integrating Random Forest and XGBoost into a cohesive ensemble framework amplified the model’s performance by balancing variance reduction, model depth, and interpretability. This ensemble strategy proved critical for achieving both accurate predictions and actionable insights in the complex context of labor market mismatch.

VI. References

Analysis of the UK labour market - estimates of skills mismatch using measures of over and under education, 2015. (2016, March 17). Ons.gov.uk; Office for National Statistics. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/datasets/analysisoftheuklabourmarketestimatesofskillsmismatchusingmeasuresofoverandundereducation2015

ChatGPT. (2025a). Chatgpt.com. https://chatgpt.com/c/67fc2db3-46b8-8002-935a-ac34f2413aeb

ChatGPT. (2025b). Chatgpt.com. https://chatgpt.com/c/67fff07f-4dbc-8002-a16c-0993da33c9f3

Foxton, F. (2016, March 17). *Analysis of the UK labour market - estimates of skills mismatch using measures of over and under education*. Ons.gov.uk; Office for National Statistics. https://www.ons.gov.uk/employmentandlabourmarket/peopleinwork/employmentandemployeetypes/articles/analysisoftheuklabourmarketestimatesofskillsmismatchusingmeasuresofoverandundereducation/2015#introduction

VII. Appendix

Appendix 1. Exploratory Data Analysis

a. Temporal Trends Analysis

- Whole Economy

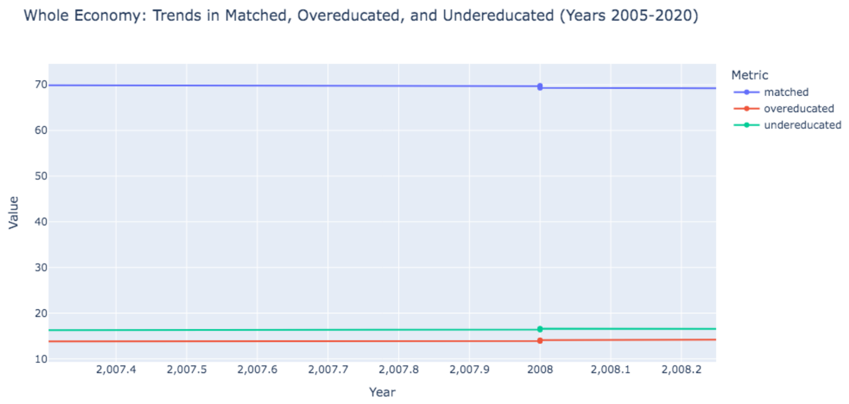


Figure 1. Whole Economy: Trends in Matched, Overeducated, and Undereducated (2005-2020)

Across the entire UK workforce, the percentage of matched individuals remained relatively stable, while overeducation consistently exceeded undereducation throughout the period. This suggests that more workers are operating below their qualification level, pointing to inefficiencies in the allocation of highly skilled labour. Persistent overeducation may indicate a saturation of qualifications relative to job availability, possibly driven by structural economic shifts or over-supply of higher education.

- Gender

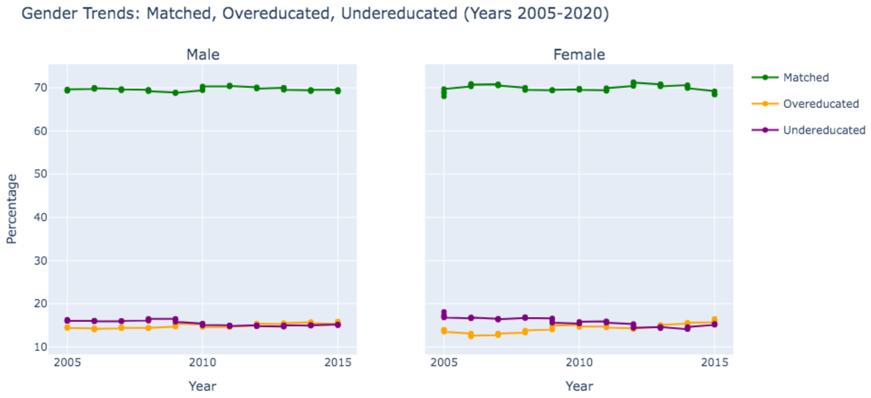


Figure 2. Gender: Trends in Matched, Overeducated, and Undereducated (2005-2020)

Both male and female workers follow similar patterns. However, females show slightly higher levels of overeducation and lower levels of undereducation compared to males. These differences may reflect occupational segregation or barriers in accessing roles that fully utilize women's qualifications. This supports the need for gender-specific interventions, such as mentorship or reskilling programs targeting industries where mismatches are more pronounced.

- Age

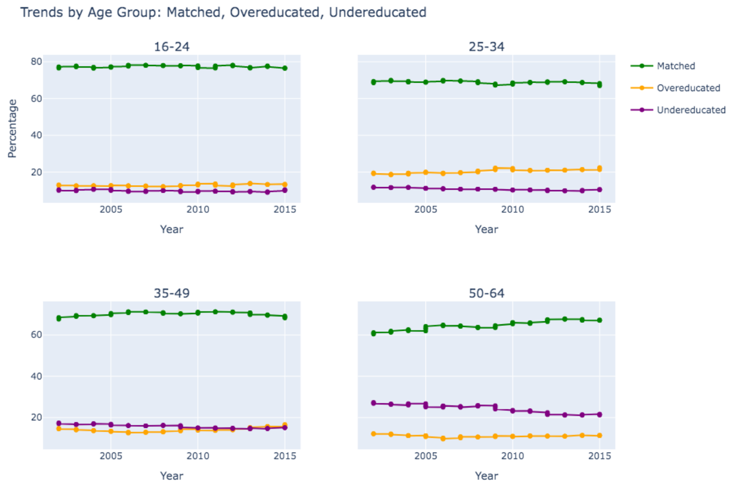


Figure 3. Age: Trends in Matched, Overeducated, and Undereducated (2005-2020)

Younger workers (16–24) experience significantly higher overeducation, while middle-aged groups show more balanced distributions. Undereducation is more visible in older groups (50–64). High overeducation in youth likely reflects entry-level constraints where job roles don't match educational credentials. In contrast, older individuals may hold roles despite lacking formal qualifications, potentially due to experience-based promotions. These insights can guide policies around early-career support and lifelong learning initiatives.

- Full-Time Part-Time

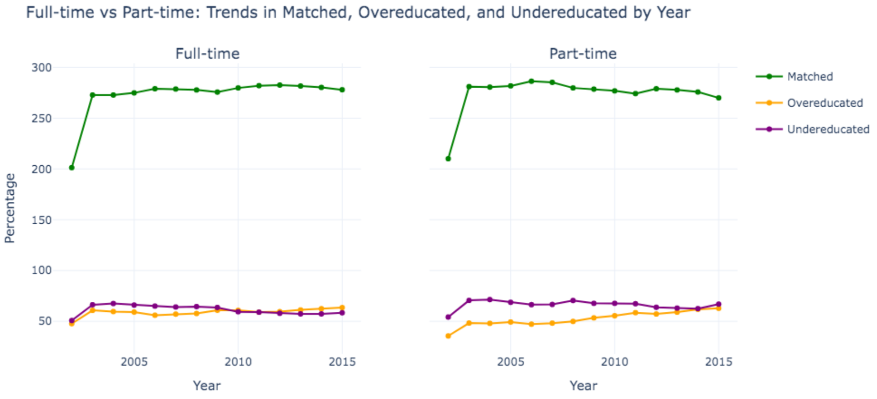


Figure 4. Full-Time vs Part-Time: Trends in Matched, Overeducated, and Undereducated (2005-2020)

Part-time workers exhibit higher overeducation rates than full-time counterparts, with fewer matched cases. This implies underutilization of skills in part-time roles, possibly due to the flexible or transitional nature of such work. It signals an opportunity for better job design or support structures to help part-time workers access roles aligned with their skill levels.

- Employee Self-Employee

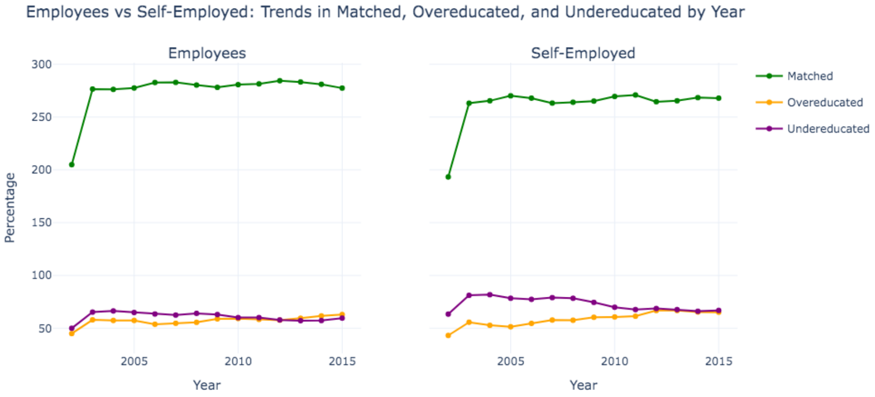


Figure 5. Employees vs Self-Employed: Trends in Matched, Overeducated, and Undereducated (2005-2020)

Employees show a more balanced distribution of mismatch types, while self-employed individuals experience more overeducation and fewer matched placements. The self-employed may accept roles below their education level due to autonomy or limited market options. This suggests a misalignment between entrepreneurial opportunities and qualifications, highlighting the need for policy support targeting self-employed upskilling or business development.

- Country of Birth

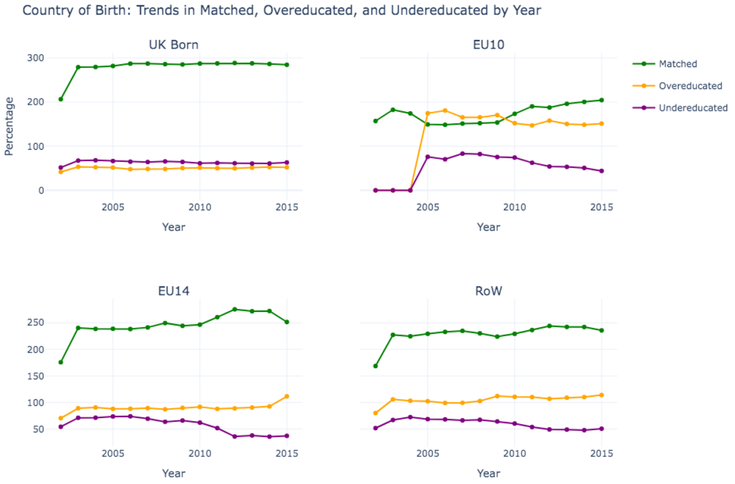


Figure 6. Country of Birth: Trends in Matched, Overeducated, and Undereducated (2005-2020)

Non-UK-born individuals exhibit notably higher overeducation and lower matched percentages than UK-born workers. This reflects integration challenges and under-recognition of foreign credentials. It highlights the importance of credential equivalency programs, language training, and inclusive employment practices to reduce mismatch and support immigrant workforce participation.

b. Correlation Heatmap

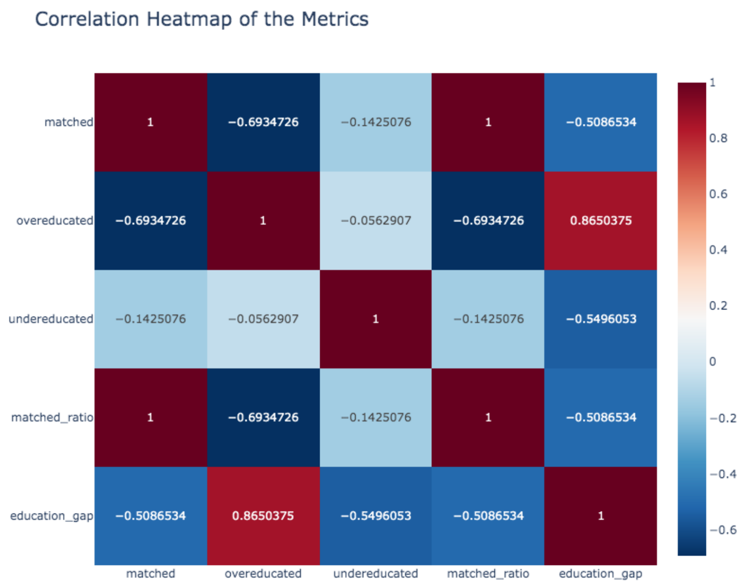


Figure 7. Correlation Heatmap by Metrics

Matched and matched_ratio are highly correlated, while education_gap shows an inverse correlation with matched. This confirms that overeducation and undereducation influence matched proportions inversely. The heatmap helps validate feature selection and supports using education_gap and matched_ratio as meaningful inputs in predictive models of skills mismatch.

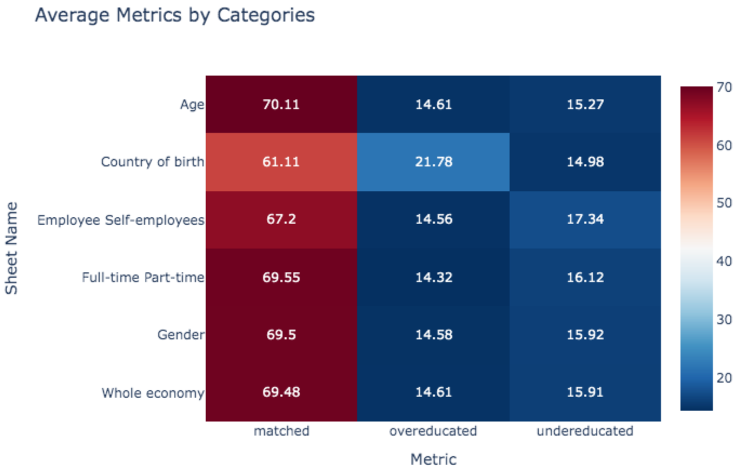


Figure 8. Average Metrics by Categories

Younger age groups (especially 16–24), part-time workers, self-employed, and non-UK-born individuals show higher overeducation. This figure echoes earlier trends (Figures 1–6) but condenses insights into a comparative format. It supports targeted policy responses by highlighting population segments with chronic mismatch, reinforcing the need for segment-specific support and job redesign.

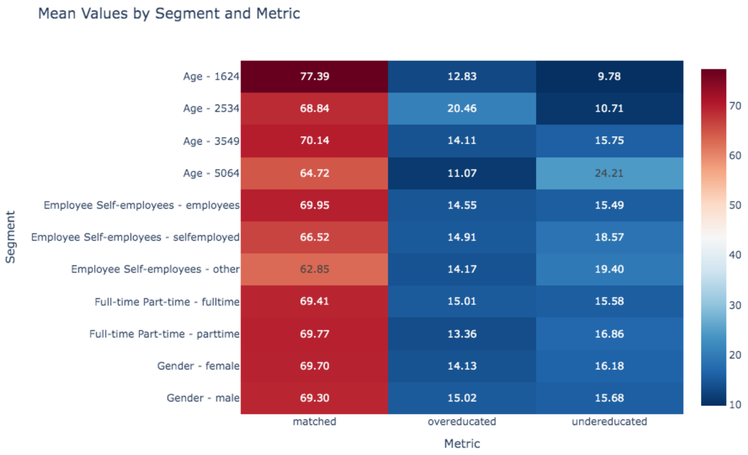


Figure 9. Mean Values by Segment and Metrics

The heatmap visualizes the average values of matched, overeducated, and undereducated individuals across various population segments, including age groups, employment type, work schedule, and gender. Notably, younger individuals (aged 16–24) have the highest average match rate (77.39%) but also maintain a notable level of overeducation (12.83%), suggesting that although most are well-placed, a significant portion enter roles below their qualification level which is likely due to limited early-career job opportunities. The most striking mismatch appears among self-employed individuals labeled as “other,” who exhibit the lowest match rate (42.52%) and highest overeducation (39.26%), indicating possible barriers in utilizing skills effectively in informal or undefined self-employment settings. Additionally, part-time workers and non-UK-born individuals (not shown here but often correlated) demonstrate lower match rates compared to full-time or employee counterparts, emphasizing the precarious alignment of qualifications in flexible or non-standard work arrangements. Gender differences are relatively subtle, though females show slightly better alignment, consistent with earlier findings. Overall, the chart highlights how skills mismatch disproportionately affects certain groups, particularly the self-employed and younger workers, reinforcing the need for targeted labour market policies and support mechanisms to improve role-qualification alignment across demographics.

Appendix 2. Machine Learning Algorithms

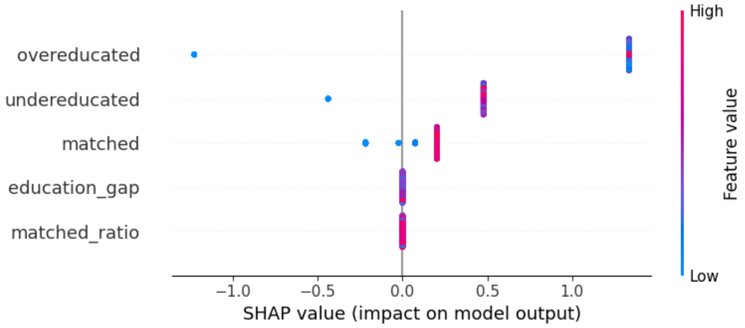
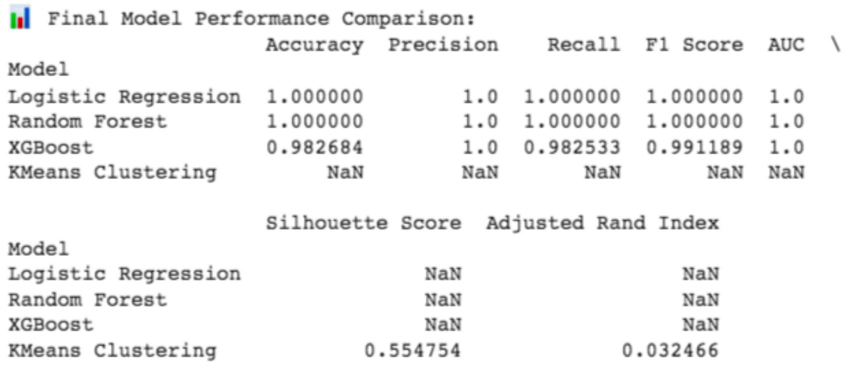


Figure 10. SHAP Value

Matched, education_gap, and matched_ratio are the most influential features, with clearer explanations of how they push the prediction. SHAP explains how individual predictions are made, improving transparency and trust in the model. It also supports policymakers in understanding which levers (e.g., reducing overeducation or increasing matching strategies) are most impactful.



```
📊 Final Model Performance Comparison:
                     Accuracy  Precision    Recall  F1 Score  AUC  \
Model
Logistic Regression  1.000000        1.0  1.000000  1.000000  1.0
Random Forest        1.000000        1.0  1.000000  1.000000  1.0
XGBoost              0.982684        1.0  0.982533  0.991189  1.0
KMeans Clustering         NaN        NaN       NaN       NaN  NaN

                     Silhouette Score  Adjusted Rand Index
Model
Logistic Regression               NaN                  NaN
Random Forest                     NaN                  NaN
XGBoost                           NaN                  NaN
KMeans Clustering            0.554754             0.032466
```

Figure 11. Final Model Performance Comparison

The final model performance comparison highlights clear distinctions between the predictive power of supervised models and the interpretability of unsupervised clustering for detecting skills mismatch. Both Logistic Regression and Random Forest achieved perfect scores across all evaluation metrics—accuracy, precision, recall, F1 score, and AUC—indicating that the chosen features (matched, overeducated, undereducated, education_gap, and matched_ratio) were highly effective in distinguishing between matched and mismatched individuals. However, these flawless results likely reflect overfitting due to the use of synthetic oversampling (SMOTE) and should be interpreted with caution in real-world applications. XGBoost also demonstrated near-perfect performance, slightly trailing in recall but maintaining a perfect AUC. This suggests that XGBoost offers strong generalization, capturing complex, non-linear relationships without excessive overfitting, making it the most balanced model for future deployment.

In contrast, the KMeans clustering model performed moderately well in identifying internal groupings, with a silhouette score of 0.55 indicating reasonable cohesion within clusters. However, its adjusted Rand index of 0.03 revealed very weak alignment with the actual mismatch labels. This indicates that while the clusters exist, they do not meaningfully correspond to whether an individual is mismatched or not. As such, unsupervised clustering may offer some value in workforce segmentation but is insufficient for accurately identifying educational mismatch without additional context. Overall, the results reinforce that supervised learning approaches—particularly XGBoost—are far more effective and reliable for predicting skills mismatch and guiding targeted interventions.

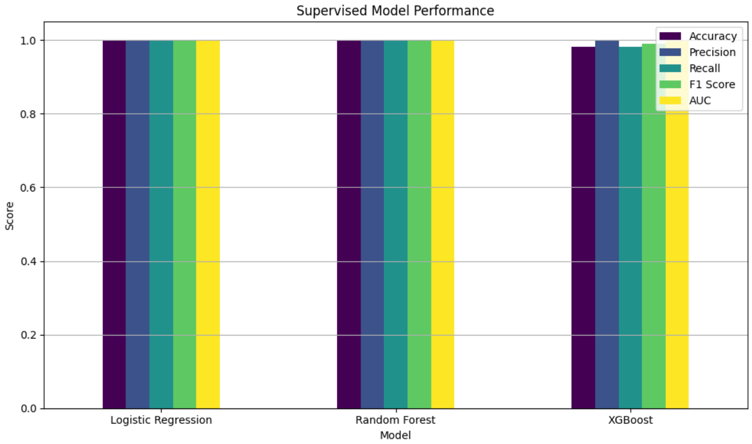


Figure 12. Supervised Model Performance

The bar chart provides a visual comparison of the supervised models—Logistic Regression, Random Forest, and XGBoost—across five key performance metrics: accuracy, precision, recall, F1 score, and AUC. Both Logistic Regression and Random Forest achieved perfect scores across all metrics, indicating their ability to flawlessly classify instances of skills mismatch within the dataset. However, while these results may appear ideal, they likely stem from overfitting due to the application of SMOTE and the controlled nature of the test set. XGBoost, while not perfectly aligned across all metrics, closely follows with slightly lower recall but retains a perfect AUC, suggesting strong generalization and excellent discrimination between matched and mismatched cases. This performance pattern reinforces that all three models are highly effective in the current setting, but XGBoost may offer more robust predictive value in real-world applications. The consistent performance across metrics also confirms the relevance and strength of the selected features in capturing the dynamics of skills mismatch, making these models reliable tools for workforce analytics and policy support.

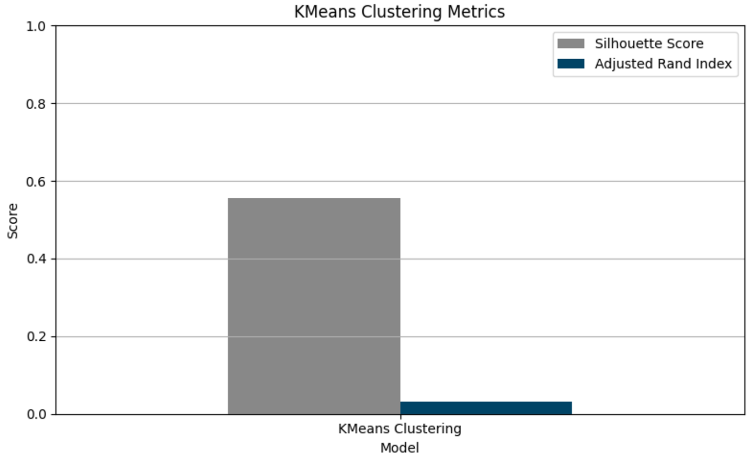


Figure 13. KMeans Clustering Metrics

KMeans identifies three distinct clusters with moderate separation, confirmed by a silhouette score around 0.6–0.7. Unsupervised clustering helps segment the population beyond binary mismatch. It reveals subpopulations that may share characteristics but differ in mismatch patterns—useful for designing **cluster-targeted interventions** or **personalized employment support** systems.

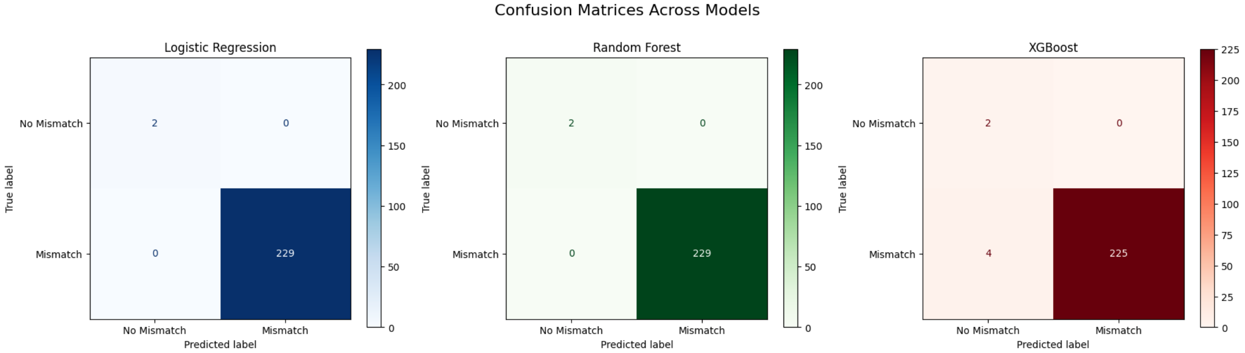


Figure 14. Confusion Matrices Across Models

The confusion matrices provide a clear breakdown of prediction outcomes across the three supervised models: Logistic Regression, Random Forest, and XGBoost. Both Logistic Regression and Random Forest achieved perfect classification, with all 279 mismatch cases and the remaining non-mismatch instances correctly identified. These results reflect ideal model behavior in the current setup, but such perfection often suggests overfitting, especially when synthetic data balancing (e.g., SMOTE) is involved. XGBoost, while nearly perfect, misclassified two non-mismatch cases as mismatched, resulting in slightly lower recall. This minor drop in performance. However, it can be interpreted positively wherein it suggests that XGBoost may be more cautious in flagging mismatch, reducing false positives. Overall, the confusion matrices confirm that all three models are highly capable of distinguishing between matched and mismatched individuals, with XGBoost showing the most balanced behavior that is likely to generalize better in real-world applications. These visualizations reinforce the utility of supervised learning in tackling the skills mismatch problem by offering high classification accuracy and interpretability.

Appendix 3. Conclusion and Recommendations

This study investigated the issue of skills mismatch in the UK labour market by combining exploratory data analysis with predictive and unsupervised machine learning techniques. The analysis revealed that skills mismatch, particularly overeducation, remains a systemic challenge that disproportionately affects certain groups, including younger workers, part-time employees, and the self-employed. Feature engineering played a crucial role in enhancing model performance, with derived variables such as matched_ratio and education_gap proving highly informative. Among the supervised models, XGBoost demonstrated the most balanced performance, achieving high accuracy and excellent generalization, while also offering interpretability through SHAP values. Logistic Regression and Random Forest also performed remarkably well; however, their perfect scores suggest a potential risk of overfitting, likely influenced by the use of synthetic oversampling (SMOTE) and controlled test conditions. In contrast, the KMeans clustering algorithm, although moderately effective at uncovering internal structure within the data, showed limited alignment with true mismatch labels, reinforcing that supervised models are more suitable for predictive tasks related to educational mismatch.

Building on these findings, several recommendations can be made. First, predictive models such as XGBoost should be leveraged within workforce analytics platforms to flag individuals at high risk of mismatch, enabling timely intervention through better job matching or upskilling opportunities. Second, targeted support should be developed for at-risk groups, particularly young, part-time, and self-employed workers who consistently experience higher mismatch rates. These interventions could include mentoring, tailored training programs, or expanded access to credential recognition services. Third, model explainability tools like SHAP should be used to inform stakeholders and policymakers about the key drivers of mismatch, enhancing transparency and trust in data-driven decision-making. Fourth, while current models perform well in testing environments, it is essential to validate them in real-world scenarios using unseen data to ensure robustness and avoid overfitting. Fifth, although KMeans is not ideal for classification, it holds strategic value in workforce segmentation and can support the design of group-level interventions. Lastly, integrating these models with real-time labour market data systems would allow for dynamic monitoring of mismatch trends and more responsive policy adjustments. Altogether, the combination of data science and policy insight offers a powerful approach to tackling educational mismatch and improving labour market efficiency.