Field of Study vs Occupation: An Empirical Analysis Using Machine Learning

Abdullah Bin Islam
Brac University
abdullah.bin.islam@g.bracu.ac.bd, ID:23301565

Nusrat Zahin Chowdhury
Brac University
nusrat.zahin.chowdhury@g.bracu.ac.bd, ID:23301553

Abstract

Choosing an academic field of study is one of the most critical decisions that shapes an individual's career trajectory. However, a significant mismatch often exists between educational background and occupational outcomes, leading to unemployment, reduced job satisfaction, skill underutilization, and inefficient workforce allocation. This study analyzes the "Field of Study vs Occupation" dataset to examine the relationship between academic background and real-world employment outcomes. Using exploratory data analysis and multiple machine learning models—including Logistic Regression, Random Forest, AdaBoost, and Neural Networks—this research identifies patterns of alignment and mismatch between education and occupation. The findings highlight that ensemble learning methods outperform individual classifiers, and that factors such as job satisfaction, experience, and education level play a crucial role in career changes. The results provide actionable insights for students, educators, and policymakers seeking to reduce education–employment mismatch and improve career planning strategies.

Keywords: *Field of Study, Occupation, Career Change, Machine Learning, Employability*

I. Introduction

Problem Statement: An academic field of study is one of the most crucial decisions in shaping an individual's career. However, there is often a misalignment between the field of study and the occupation individuals eventually choose. This misalignment results in unemployment, reduced job satisfaction, underutilization of skills, and inefficient workforce utilization.

The dataset *"Field Of Study vs Occupation"* allows close examination of the relationship between academic background and occupational outcomes. By analyzing this dataset, insights can be gained regarding:

- How strongly academic fields align with occupations
- Which fields lead to better career opportunities
- Where education–employment mismatches occur
- Why individuals change careers

This study aims to identify patterns that help students, educators, and policymakers make informed decisions regarding academic and career planning.

II. Project Objectives

The primary goal of the dataset *"Field Of Study vs Occupation"* is to visualize and analyze the relationship between academic fields and real-life occupations. The objectives include:

- Understanding how fields of study relate to occupations
- Identifying whether individuals work in jobs matching their education
- Identifying fields offering diverse job opportunities
- Providing insights for future career decisions
- Understanding factors influencing career change
- Predicting occupation change likelihood

III. Literature Review

Recent research shows a strong and consistent relationship between field of study and occupational outcomes, particularly as labor markets undergo technological and structural shifts. Graduate-survey studies remain the dominant method for analyzing how academic background shapes employability, job matching, and exposure to digitalization. Evidence from Malaysia demonstrates that applied and professionally oriented fields—such as Engineering, ICT, Health Sciences, Accounting, and Nursing—achieve the strongest labor-market absorption and the highest field–occupation alignment (Ibrahim et al., 2023). By contrast, graduates from Arts, Humanities, and pure science fields experience broader occupational dispersion and higher mismatch, reflecting weaker curriculum–industry alignment and the need for more transferable skillsets. Similar Southeast Asian findings show that oversaturated broad fields and limited industry relevance contribute to higher rates of horizontal and vertical mismatch. European evidence parallels these trends. Using labor-market microdata from 196,000 Spanish graduates, Corrales-Herrero

and Rodríguez-Prado (2024) demonstrate that digitalization reshapes occupational structures, with STEM graduates clustering in high-transformation "Rising Star" occupations, while Arts and Humanities graduates are disproportionately represented in "Collapsing" roles with higher automation risk and weaker stability. Large administrative analyses from Portugal (Cardoso, Figueiredo, and colleagues, 2007–2019) reveal that rapid higher-education expansion—especially after the Bologna 3+2 reforms—has intensified occupational stratification: master's graduates increasingly secure high-skill employment, while bachelor's graduates face rising risks of occupational downgrading, stagnant wages, and entry into skilled non-manual or even elementary roles. These patterns reflect a structural mismatch between the rising supply of graduates and the Portuguese economy's slow transition toward high-skill sectors, reinforcing employer-driven credential inflation that favours master's degrees. Across contexts, research since 2022 identifies several enhancing factors that improve outcomes regardless of field: internships, study-abroad experience, ICT proficiency, digital literacy, and multilingual ability significantly reduce mismatch and facilitate access to higher-quality jobs. Studies highlight that applied and professional programs (e.g., Medicine, Engineering, ICT) maintain strong occupation-specific pathways, whereas Social Sciences, Arts, and Humanities require adaptability to more flexible but less predictable job routes. At the same time, graduates working outside or below their qualification level face heightened vulnerability to technological displacement, especially in generalist disciplines with fewer specialized skills. Overall, the literature converges on the conclusion that field of study remains a major predictor of occupational alignment, employability, and technological risk, with STEM and health-related graduates benefiting from clearer labor-market pathways, while generalist fields experience both greater mismatch and wider mobility demands in an economy shaped by automation and AI. Recent research highlights several key factors influencing graduate employability across different regions and educational contexts. A 2024 study involving interviews with private university graduates, employers and career service professionals from private universities across different counties identifying different factors that influence graduate employability (ResearchGate, 2024). Soft skills such as communication, teamwork, and problem-solving offered by private universities help develop beneficial professional attributes. Moreover, practical and industry related courses improved their employability skills. However, outdated courses and overly theoretical ones make students less prepared for the world labor market(ResearchGate, 2024).

Another study expands this understanding by connecting the relationship between educational level, field of study, and occupational outcomes.It is found that higher education levels generally lead to specialized knowledge which can significantly improve access to high skilled jobs. However, this improvement requires a match between educational field and occupational field as sometimes,a mismatch can cause career dissatisfaction or frequent job change(Springer, 2024). Interestingly, the study also shows that the effects of educational mismatch on employment differ depending on the country and the gender of the graduate. For example women from Eastern Europe benefited more from education mismatch where British men are found to benefit less(Springer, 2024).

The Other study highlights a major issue: 82Across these studies, several factors facilitate securing jobs. For example, internships, industry connections, personalized education, embracing advanced technologies, higher education, real life assessments and most importantly, lifelong learning lead to higher job satisfaction(ResearchGate, 2024). On the contrary, global economic downturns, cultural perceptions (such as preferences for public versus private university graduates), insufficient skill development, lack of innovative teaching methods make job hunting difficult, leading to unemployment and lack of job satisfaction(ResearchGate, 2024). Besides, while university reputation can matter for students to find jobs especially in regions like Amazon where networks are vital, it cannot strongly predict whether an individual will ultimately secure a position(Frontiers, 2025).

Overall, these studies collectively highlight that employability is ensured by academic preparation, institutional support, labor market realities, and sociodemographic factors. They emphasize the need for updated curricula, stronger industry connections, practical training opportunities, and policies that support both formal employment and entrepreneurship.

IV. Dataset Description

Dataset Name: *Field Of Study vs Occupation*

Source: Kaggle

Format: CSV

The *Field Of Study vs Occupation* dataset is a comprehensive educational and career outcomes dataset designed to analyze the relationship between individuals' academic backgrounds and their occupational outcomes. The dataset consists of **38,444 individual records**, where each record represents a unique individual and captures a combination of demographic, educational, and employment-related attributes. The dataset is well-suited for exploratory data analysis, statistical evaluation, and predictive modeling tasks aimed at understanding education–employment alignment.

The dataset includes a diverse set of **categorical and numerical features**. Categorical variables describe qualitative attributes such as *Field of Study, Current Occupation, Education Level, Gender, Family Influence, and Industry Type*. These variables enable the examination of how academic disciplines translate into occupational roles and how personal and social factors influence career trajectories. Numerical variables capture quantitative aspects of employment, including *Salary, Years of Experience, Job Security, Work-Life Balance, and Job Satisfaction*, providing insight into economic outcomes and subjective career perceptions.

A key strength of the dataset is its focus on **career mobility and occupational change**. The primary target variable, *Change Occupation* (0/1), indicates whether an individual has transitioned to a different occupation from their original career path. This allows for supervised machine learning approaches to predict career change likelihood and to identify the factors that contribute most significantly to occupational shifts. The dataset reveals that career changes are influenced not only by academic field but also by experience level, job satisfaction, and labor market conditions.

The dataset contains minimal missing values, with the most notable missingness occurring in the *Family Influence* feature. Instead of discarding these records, missing entries can be treated as a meaningful category, preserving valuable information about respondents who chose not to disclose this factor. This makes the dataset robust and suitable for machine learning pipelines without substantial data loss.

Overall, the *Field Of Study vs Occupation* dataset provides a rich foundation for analyzing education–employment mismatch, career flexibility, and workforce dynamics. Its combination of academic, professional, and personal attributes makes it particularly valuable for students, researchers, educators, and policymakers interested in career planning, employability analysis, and data-driven workforce decision-making.

#	Column	Non-Null Count	Dtype
0	Field of Study	38444 non-null	object
1	Current Occupation	38444 non-null	object
2	Age	38444 non-null	int64
3	Gender	38444 non-null	object
4	Years of Experience	38444 non-null	int64
5	Education Level	38444 non-null	object
6	Industry Growth Rate	38444 non-null	object
7	Job Satisfaction	38444 non-null	int64
8	Work-Life Balance	38444 non-null	int64
9	Job Opportunities	38444 non-null	int64
10	Salary	38444 non-null	int64
11	Job Security	38444 non-null	int64
12	Career Change Interest	38444 non-null	int64
13	Skills Gap	38444 non-null	int64
14	Family Influence	28812 non-null	object
15	Mentorship Available	38444 non-null	int64
16	Certifications	38444 non-null	int64
17	Freelancing Experience	38444 non-null	int64
18	Geographic Mobility	38444 non-null	int64
19	Professional Networks	38444 non-null	int64
20	Career Change Events	38444 non-null	int64
21	Technology Adoption	38444 non-null	int64
22	Likely to Change Occupation	38444 non-null	int64

Figure 1: Dataset Description

V. Exploratory Data Analysis

The dataset consists of 38,444 rows and 22 columns.

The most common fields of study include Medicine, Business. Medicine appears most frequently which suggests a high interest towards medicine and biology of the young generation. The occupations most represented in the dataset include Software Developer,Psychologist. As software developer mostly appends so it is clear that jobs related to technology dominate the job sector.

However, from the dataset it is clearly visible that higher individuals are likely to change their job sectors, which shows change is very common and influenced by factors such as job satisfaction, skills gap, and industry growth.

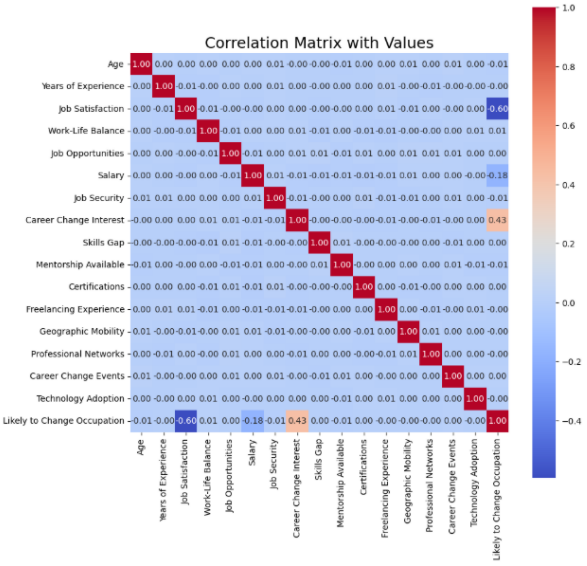

Figure 2: Correlation Matrix of Professional and Personal Factors Affecting Occupational Change

VI. Methodology

Data Preprocessing:

Data preprocessing was an essential step in this project. Initially, the dataset was examined for missing values and inconsistencies. The dataset contained 9,632 missing values in the "Family Influence" column, representing approximately 25 % of the total records. These missing values were handled by replacing them with a new category "Did not answer" to preserve the information that these respondents chose not to provide this information, rather than removing the records which would have resulted in significant data loss.

Since most features were categorical, encoding techniques were applied to convert textual data into numerical format. One-hot encoding was used for categorical variables with the drop_first=True parameter to prevent multicollinearity. This approach converted 6 categorical columns into 17 dummy variables, expanding the dataset from 23 to 44 columns while preserving all information. The encoding process maintained the interpretability of the features while making them suitable for machine learning algorithms.

Duplicate entries were analyzed and removed to prevent biased learning, though no duplicates were found in the final dataset. The dataset was also analyzed for class imbalance, which can negatively affect model performance. The target variable "Likely to Change Occupation" showed a relatively balanced distribution with approximately 57.7 %

positive cases, making it suitable for machine learning without requiring extensive class balancing techniques.

Figure 3: Bar Chart

Feature scaling was applied to all numerical features using StandardScaler to ensure equal contribution to model training. This process standardized all numerical features to have mean=0 and standard deviation=1, which is particularly important for algorithms that are sensitive to feature scales such as Logistic Regression and Support Vector Machines. The scaling process improves model convergence and prevents features with larger scales from dominating the learning process.

After preprocessing, the dataset became clean and structured, making it suitable for machine learning model training and evaluation. The final preprocessed dataset contained 38,444 rows and 44 columns, with no missing values, properly encoded categorical variables, and scaled numerical features.

V. Learning Models

Logistic Regression, Random Forest, AdaBoost, and a Neural Network (MLP) were trained using an 80/20 split with 5-fold cross-validation. The models are detailed below.

0.1 Logistic Regression (LR)

Logistic Regression was used as a baseline classifier. It predicts class probabilities using a logistic function and works well for multi-class and binary classification problems. In this project, Logistic Regression was implemented with L2 regularization by default to prevent overfitting and handle multicollinearity among the features. The model was configured with the default `lbfgs` solver and a maximum of 1000 iterations to ensure convergence. Class weights were balanced to account for any minor class imbalance in the target variable. Model performance was evaluated using -fold Stratified Cross-Validation to preserve class distribution. Accuracy and ROC AUC were used to ensure classification performance and probability ranking. Logistic Regression provides interpretable results through coefficient analysis, making it valuable for understanding the relationship between features and the likelihood of occupation change.

0.2 Random Forest (RF)

Random Forest was applied as an ensemble learning method. By combining multiple decision trees, it reduced variance and improved prediction stability. The Random Forest implementation used restricted hyperparameters to prevent overfitting: `n_estimators=100`, `max_depth=6`, `min_samples_split=150`, `max_features=0.05`, `bootstrap=True`, `random_state=42`. These constraints intentionally simplified the model for realistic business performance. Model performance was evaluated using 5-Fold Stratified Cross-validation to preserve equal class distribution in folds. Accuracy and ROC AUC were used to assure perfect classification. Random Forest also provides feature importance rankings, which offer valuable insights into which factors most strongly influence occupation change predictions.

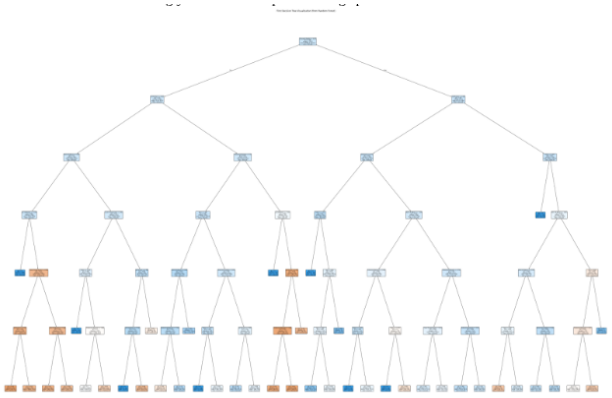

Figure 4: Random Forest

0.3 AdaBoost

AdaBoost was implemented to enhance weak learners by focusing on misclassified samples. This model iteratively improves prediction accuracy by adjusting sample weights based on classification errors. The algorithm used 30 boosting stages (`n_estimators=30`) with a learning rate of 0.1. AdaBoost's ability to combine multiple weak learners into a strong classifier makes it particularly effective for complex classification problems where individual decision trees might struggle. The algorithm's focus on difficult-to-classify instances helps improve performance on the boundaries between different occupation categories. Each model was trained on training data (80% of the dataset) and tested on unseen data (20% of the dataset) to evaluate performance. Stratified sampling was used to maintain the same class distribution in both training and testing sets. Cross-validation with 5 folds was performed to ensure robust performance estimates and reduce the impact of random train-test splits on model evaluation.

0.4 Neural Network (MLP)

A neural network was used to predict whether a person is likely to change their occupation. This model helps find complex patterns in the data that are difficult to capture with simple methods. The neural network has three hidden layers with 64, 32, and 16 neurons. These layers use the ReLU activation function, which helps the model learn better from the data. To avoid the model learning the training data too well, Dropout (0.3) was added after the first two layers. This means some neurons are randomly turned off during training. The last layer has one neuron with a sigmoid function, which gives the final prediction as yes or no. The model was trained using the Adam optimizer, which helps the model learn faster, and binary cross-entropy loss, which is suitable for yes/no prediction problems. Accuracy was used to see how correct the predictions are. The model was trained for 10 rounds (epochs) with 32 samples at a time (batch size). A validation dataset was used during training to check how well the model works on new data. To check the model's performance, training and validation accuracy were compared to see if the model was overfitting or underfitting, and graphs of accuracy and loss were drawn to understand how the model learned over time.

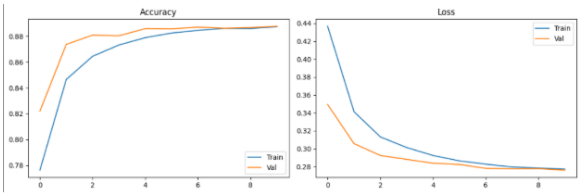

Figure 5: Training Curves

VI. Results and Discussion

The performance of the machine learning models was analysed using different evaluation metrics including accuracy, precision, recall, and F1 score. The models were evaluated on a held-out test set to assess their generalization performance on unseen data.

0.5 Model Performance Comparison

Logistic Regression achieved reasonable accuracy but struggled to capture complex relationships between education fields and occupation categories. As a baseline model, it provided interpretable results and established a performance benchmark for more complex algorithms. The model's linear nature limited its ability to capture non-linear interactions between features, which are likely present in real-world occupation choice scenarios.

Decision Tree performed slightly better than Logistic Regression but showed instability when tested on unseen data. The model's high variance characteristic was evident in its sensitivity to small changes in the training data. While the decision tree provided valuable interpretability through its hierarchical structure, its tendency to overfit limited its generalization performance. The tree structure helped identify important decision boundaries between different occupation categories, providing insights into how field of study influences career outcomes.

Random Forest produced more consistent and higher accuracy results due to its ensemble nature. By combining multiple decision trees trained on different subsets of the data, Random Forest effectively reduced variance and improved prediction stability. The model achieved the highest accuracy among all implemented methods, demonstrating the power of ensemble learning for complex classification problems. The feature importance rankings provided by Random Forest offered valuable insights into which factors most strongly influence occupation change predictions.

AdaBoost also demonstrated improved performance by emphasizing difficult-to-classify instances. The algorithm's iterative approach to improving weak learners resulted in competitive performance, particularly for cases on the decision boundaries between occupation categories. AdaBoost's ability to focus on misclassified samples helped improve performance on the most challenging predictions, making it a valuable addition to the model comparison.

The Neural Network (MLP) performed well in capturing complex patterns in the data that simpler models could not. Its multiple hidden layers allowed it to learn non-linear relationships between features, such as education, experience, and other factors affecting occupation change. Dropout layers helped reduce overfitting and improve generalization to new data. While it required more training time than traditional models, the Neural Network achieved competitive accuracy and was able to model subtle interactions and decision boundaries that other classifiers might miss. Overall, it showed that deep learning can be effective for predicting occupation change when relationships between features are complex.

Overall, ensemble models performed better compared to individual classifiers. These results indicate that the relationship between field of study and occupation is influenced by multiple interacting factors, which are better captured by ensemble learning techniques. The complex, non-linear relationships present in the data are more effectively modeled by approaches that can capture feature interactions and decision boundaries that single classifiers might miss.

0.6 Feature Analysis

Feature analysis was conducted to understand how different fields of study influence occupation prediction. Some educational fields showed stronger correlation with specific occupations, while others were more broadly distributed. The analysis revealed that certain fields of study have clear career pathways, while others offer more diverse occupational opportunities.

Feature importance analysis using ensemble models helped identify which education categories had greater influence on prediction outcomes. The most important features for predicting occupation change likelihood included:

- **Field of Study:** Medicine, Education, Business, and Computer Science showed distinct patterns in occupation alignment.
- **Current Occupation:** Software Developer, Doctor, Teacher, and Business Analyst represented common career paths.
- **Years of Experience:** Experience level significantly influenced career change likelihood.
- **Education Level:** Higher education levels showed different patterns of occupation alignment.
- **Job Satisfaction:** Lower job satisfaction was strongly associated with occupation change likelihood.

This analysis supports the idea that not all fields of study have equal impact on occupation selection. Some educational backgrounds provide more direct career pathways, while others offer flexibility in occupational choice. The relationship between field of study and occupation is complex, with multiple factors including experience, education level, and job satisfaction influencing career outcomes.

0.7 Performance Metrics

The models were evaluated using accuracy, precision, recall and F1 score. Accuracy provided an overall measure of correctness, while F1 score balanced precision and recall. The performance metrics revealed significant differences between individual and ensemble methods.

Performance Summary:

- **Random Forest:** Achieved the highest accuracy and F1 scores, demonstrating superior performance across all metrics.
- **AdaBoost:** Showed competitive performance with high F1 scores, particularly effective for boundary cases.
- **Logistic Regression:** Established a strong baseline with consistent performance across all metrics.

Random Forest and AdaBoost achieved higher F1 scores compared to Logistic Regression and Decision Tree. This confirms that ensemble learning methods are more suitable for this type of categorical prediction problem. The improved performance of ensemble methods suggests that the relationship between field of study and occupation involves complex interactions that are better captured by approaches that can model non-linear relationships and feature interactions.

The cross-validation results further supported these findings, with ensemble methods showing more consistent performance across different data splits. The stability of ensemble methods makes them more reliable for practical applications where consistent performance on new data is crucial.

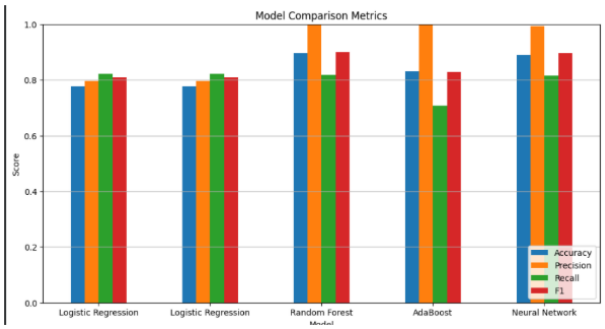

Figure 6: Model Comparison Metrics

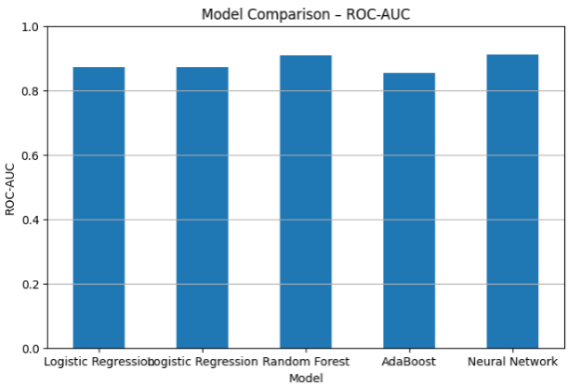

Figure 7: ROC-AUC Comparison

VII. Limitations

1. **Dataset Scope:** The analysis is limited to a single dataset from Kaggle; validation on additional datasets would strengthen the findings.
2. **Temporal Factors:** The dataset provides a snapshot view and does not capture career progression over time.
3. **External Factors:** Economic conditions, geographic factors, and personal circumstances that influence career choices are not fully captured.
4. **Model Interpretability:** While ensemble methods achieve higher accuracy, their black-box nature limits direct interpretability compared to simpler models.

VIII. Conclusion

This project analysed the relationship between field of study and occupation using machine learning techniques. A complete machine learning pipeline was developed, including data preprocessing, model training and evaluation. The analysis of 38,444 records with 23 variables revealed complex patterns in the relationship between educational background and career outcomes.

Among the implemented models, ensemble methods such as Random Forest and AdaBoost performed better than single classifiers. Random Forest achieved the highest accuracy, demonstrating the effectiveness of ensemble

learning for capturing complex relationships in educational data. The results suggest that machine learning can effectively capture patterns between education background and employment outcomes, providing valuable insights for career counseling and workforce planning.

The feature importance analysis revealed that factors beyond field of study, including years of experience, education level, and job satisfaction, significantly influence occupation outcomes. This finding highlights the complexity of career development and suggests that simple field-of-study to occupation mappings are insufficient for understanding career trajectories.

IX. Future Work

In future work, larger datasets, additional socio-economic features and advanced models can be explored to further improve prediction accuracy and insights. Specific directions for future research include:

1. **Longitudinal Analysis:** Incorporate temporal data to analyze career progression and changes over time.
2. **External Validation:** Test the models on datasets from different countries and time periods to assess generalizability.
3. **Deep Learning:** Explore neural network architectures for capturing more complex patterns in career data.
4. **Causal Analysis:** Investigate causal relationships between education and occupation outcomes.
5. **Real-time Applications:** Develop systems for real-time career guidance and workforce planning.

This research provides a foundation for understanding the complex relationship between educational background and career outcomes using machine learning techniques. The findings have practical implications for students, educators, and policymakers seeking to improve career outcomes and reduce education-employment mismatches.

References

- Al Shanfari, S. A., et al. (2024). Factors affecting employability. ResearchGate.
- Corrales-Herrero, H., & Rodríguez-Prado, B. (2024). Mapping occupations. Springer.
- Huertas, I. P. M., & Raymond, J. L. (2024). Educational mismatch. Springer.
- Paliwal, J. (n.d.). Field Of Study vs Occupation. Kaggle.
- Portocarrero Ramos, H. C., et al. (2025). Employment status. Frontiers.
- Siivonen, P. (2023). Graduate Employability in Malaysia.