analytics

MDPI

Article

Unveiling the Impact of Socioeconomic and Demographic Factors on Graduate Salaries: A Machine Learning Explanatory Analytical Approach Using Higher Education Statistical Agency Data

Bassey Henshaw [1], Bhupesh Kumar Mishra [2,*], William Sayers [1] and Zeeshan Pervez [3]

[1] School of Computing and Technology, University of Gloucestershire, Cheltenham GL50 2RH, UK; henshaw49@gmail.com (B.H.)
[2] Centre of Excellence for Data Science, Artificial Intelligence and Modelling (DAIM), University of Hull, Cottingham Road, Hull HU6 7RX, UK
[3] School of Engineering, Computing, and Mathematical Sciences, University of Wolverhampton, Wolverhamton WV1 1LY, UK
* Correspondence: bhupesh.mishra@hull.ac.uk

Abstract: Graduate salaries are a significant concern for graduates, employers, and policymakers, as various factors influence them. This study investigates determinants of graduate salaries in the UK, utilising survey data from HESA (Higher Education Statistical Agency) and integrating advanced machine learning (ML) explanatory techniques with statistical analytical methodologies. By employing multi-stage analyses alongside machine learning models such as decision trees, random forests and the explainability with SHAP stands for (Shapley Additive exPanations), this study investigates the influence of 21 socioeconomic and demographic variables on graduate salary outcomes. Key variables, including institutional reputation, age at graduation, socioeconomic classification, job qualification requirements, and domicile, emerged as critical determinants, with institutional reputation proving the most significant. Among ML methods, the decision tree achieved a standout with the highest accuracy through rigorous optimisation techniques, including oversampling and undersampling. SHAP highlighted the top 12 influential variables, providing actionable insights into the interplay between individual and systemic factors. Furthermore, the statistical analysis using ANOVA (Analysis of Variance) validated the significance of these variables, revealing intricate interactions that shape graduate salary dynamics. Additionally, domain experts' opinions are also analysed to authenticate the findings. This research makes a unique contribution by combining qualitative contextual analysis with quantitative methodologies, machine learning explainability and domain experts' views on addressing gaps in the existing identification of graduate salary predicting components. Additionally, the findings inform policy and educational interventions to reduce wage inequalities and promote equitable career opportunities. Despite limitations, such as the UK-specific dataset and the focus on socioeconomic and demographic variables, this study lays a robust foundation for future research in predictive modelling and graduate outcomes.

Keywords: graduate salaries; higher education; machine learning; socioeconomic and demographic factors; statistical analysis; SHAP; analysis of variance (ANOVA)

Academic Editor: Carson K. Leung

Received: 2 January 2025
Revised: 18 February 2025
Accepted: 4 March 2025
Published: 11 March 2025

Citation: Henshaw, B.; Mishra, B.K.; Sayers, W.; Pervez, Z. Unveiling the Impact of Socioeconomic and Demographic Factors on Graduate Salaries: A Machine Learning Explanatory Analytical Approach Using Higher Education Statistical Agency Data. *Analytics* **2025**, *4*, 10. https://doi.org/10.3390/analytics4010010

1. Introduction

The expansion of higher education in the past few decades has made baccalaureate degrees more accessible to students [1]. However, universities have historically perpetuated social inequality by being elitist institutions that only admit a select few. This began to change when higher education started to become more democratic and open [2]. In today's world, universities are generally expected to train a part of the active population, continuously review their offers, and adapt them to current market demands and future professional requirements [3]. Universities are also taking responsibility for integrating graduates into the labour market [4]. Graduates' entry into the labour market is a critical mechanism through which public investment in higher education reveals its returns and the returns to investment in human capital depend on the use that graduates can make of their education. It is no surprise that the global higher education sector remains vibrant as more and more students apply to study at universities, given that graduate incomes are significantly higher than non-graduates [5], and higher education brings many benefits to students and society, driving economic growth [6].

However, despite the expansion of higher education, disparities in graduate salaries persist, influenced by a complex set of factors. Socioeconomic background, institutional reputation, and demographic variables all contribute to wage differences, yet their combined effects remain unclear. While previous studies have explored some of these factors individually, there is limited research on their collective impact using advanced data-driven techniques. Furthermore, the existing literature presents conflicting findings, particularly regarding the role of the socioeconomic status in salary progression. Some research suggests that graduates from lower-income backgrounds face long-term pay disadvantages, while other studies indicate that factors like institutional prestige and job qualifications outweigh socioeconomic influences. This inconsistency highlights the need for a comprehensive, data-driven analysis to disentangle the interactions between these factors and their impact on graduate earnings.

The purpose of this research is to analytically investigate the relationship between socioeconomic and demographic factors and graduate salaries using machine learning algorithms. Understanding this relationship is crucial for informing policymaking, operational planning, and financing in higher education institutions. For instance, by identifying the factors that significantly impact graduate salaries, universities can tailor their programmes and support services to better prepare students for the job market. Additionally, knowledge of these factors can help policymakers design interventions that promote equal access to education and career opportunities, ultimately contributing to economic and social development.

This study addresses a critical gap by integrating machine learning with traditional statistical techniques to provide a more precise, data-driven understanding of salary determinants. Unlike previous works that rely on conventional regression models, this research leverages advanced ML explainability methods such as SHAP (Shapley Additive Explanations) to uncover complex relationships between graduate salaries and key socioeconomic variables. By using a large-scale dataset from the Higher Education Statistical Agency (HESA), this study offers a robust and scalable approach to identifying the most influential salary determinants, thereby enhancing the predictive power of salary estimation models. This study further aims to contribute to the existing body of knowledge by providing a comprehensive analysis of the determinants of graduate salaries, which can serve as a foundation for evidence-based decision-making in the higher education sector. Furthermore, this study aims to build on existing studies that have identified factors such as class of degree, fields of study, type of university attended, and the age of the graduate as contributing factors to graduate salaries [7]. However, there are contradictory findings in the literature,

such as the relationship between the socioeconomic status and pay progression [8], which require further investigation.

1.1. Global Higher Education (HE) Growth

Higher Education (HE), also known as post-secondary education, third-level, or tertiary education, is an optional final stage of formal learning that occurs after the completion of secondary education. The globalisation of higher education has presented significant challenges to traditional concepts of a university, which has historically been viewed as the citadel of elite liberal education. However, today, universities are seen as accessible sources of relevant and applicable information across the globe [9]. While this transition has unwanted and unexpected consequences, there is a growing debate about whether HE is worthwhile and whether it generates social mobility as more and more money is spent on it internationally [10].

Between 1962 and 1998, university education in the UK was paid for through taxes, and enrolment rates increased significantly during that time, rising from 7% of all high school graduates in 1962 to more than 30% by the late 1990s [11]. Several successive governments, aware of the economic benefits of further growth in higher education, but alarmed by increasing costs, began shifting the cost from the government onto students in 1998 when tuition fees were introduced across the UK at GBP 1000 per year. In summary, the maximum fee in England for students domiciled in England, Northern Ireland, and Wales increased dramatically from 2006 to 2012, when it was raised from GBP 3000 to GBP 9000, later increased to GBP 9250 in 2017/18 and will be GBP 9535 from the year 2025/26. Northern Irish students at Northern Irish HEIs experienced a much smaller increase with fees capped at GBP 3805 since 2012. Welsh students were also subject to modest fee increases, capped at GBP 4045. The only country to offer free university education during this period was Scotland [12].

The university has indeed undergone a significant revolution, incorporating training for labour insertion among its objectives for graduates, which, in turn, increases graduate satisfaction, as identified and measured by the amount of salary earned by a prospective graduate. Academic accreditation is perceived to improve one's career prospects and generate financial profits, which is one of the primary factors attracting students [13]. However, despite this expectation, the reality appears to be more ambiguous. However, several determinants play a role in influencing these graduates' aspirations, including socioeconomic, demographic, and cultural factors. In the realm of academia, where aspirations are nurtured, a pressing concern lingers—wage inequality among graduates possessing comparable qualifications within the same country and field of study.

This pervasive imbalance has the potential to impede career progression and limit earning prospects. Individuals from underprivileged backgrounds may face systemic barriers in accessing quality education and professional networks, which can hinder their ability to secure high-paying jobs or advance in their careers [14]. Additionally, unconscious bias and discriminatory practices in recruitment and promotion processes may further exacerbate this imbalance, perpetuating disparities in earning potential. Addressing these systemic issues is crucial for promoting fairness and equity in the labour market, ensuring that all individuals have equal opportunities to succeed and thrive in their careers regardless of their socioeconomic backgrounds.

Furthermore, existing studies in this field have provided valuable insights into wage inequality, shedding light on factors such as cost of living, course specialisation, level of qualification, employment sector, and demographic dynamics [15]. However, these studies often lack a comprehensive analysis of the interplay between these factors and may not account for evolving socioeconomic and cultural contexts. The role of unconscious bias

and systemic discrimination in wage disparities has been increasingly recognised in recent studies. For instance, Bertrand and Duflo [16] argue that unconscious bias continues to influence hiring and pay practices, even in industries striving for equality. They also note that such biases often intersect with systemic barriers, including unequal access to education and career advancement opportunities. Similarly, Blau and Kahn [17] highlight the persistence of gender and racial wage gaps, which cannot be fully explained by observable factors like education or experience. Smith et al. [18] explore how microaggressions and other subtle forms of bias impact workplace dynamics and contribute to wage inequities, especially for marginalised groups. Additionally, Chetty et al. [19] underscore the influence of systemic factors, such as geographic segregation and differential access to networks, on income mobility and wage outcomes.

By incorporating a multifactorial analysis, this study seeks to build on these foundational works. It integrates data-driven insights and interpretability methods SHAP to untangle the complex interactions between demographic, educational, and systemic variables in wage inequality, offering a more nuanced understanding of these disparities. This approach directly addresses limitations noted by previous studies, such as their reliance on broad aggregates that obscure finer patterns of discrimination [20]. In addition, the geographical span of this study includes the entire United Kingdom, dissecting the socioeconomic and demographic dimensions that intersect with graduate salaries and labour market assimilation. While the focus of this study is on the UK due to data availability, the findings and analysis can provide valuable insights for policymakers, educational institutions, and researchers in other countries/regions.

By understanding the complex relationship between socioeconomic and demographic factors and graduates' salaries, this work can inform the development of policies, financing strategies, and educational programmes aimed at promoting equal access to education and career opportunities worldwide. With the use of qualitative and quantitative data analysis at its disposal, sourced from the merged HESA and Jisc, this study unravels the impact of post-graduation credentials, degree classifications, and the commitment invested in the chosen course of study upon job attainment and associated compensation. It is within this landscape that machine learning techniques expand, casting illumination upon the path towards well-informed employment decisions and seamless integration into the competitive job market. Also, by scrutinising the socioeconomic and demographic facets that intricately interlace with graduate salaries in the UK, this study unearths the bedrock of this disparity. The introduction of rule-based patterns and the pivotal role of SHAP magnify this study's significance. The application of ML/AI to graduate survey data adds an innovative section to the literature on Higher Education Data Analytics, accelerating knowledge discovery and amplifying the realms of explainable artificial intelligence.

In this illumination, the analytical finding highlights the direction towards rectifying the imbalances, thereby augmenting the prospects of graduates and fostering a more equitable higher education landscape. In other words, this research embarks on a comprehensive exploration of the intricate factors influencing graduate salaries, addressing key questions to uncover the nuanced dynamics at play. It stands out by focusing on socioeconomic and demographic variables that significantly influence graduate salaries and evaluating how different modelling techniques comprehend these factors. Furthermore, it examines potential synergies between these contributors and their cumulative impact, a relatively underexplored area in the existing literature.

1.2. Paper Organisation

The remainder of this paper is structured as follows: Section 2 presents a comprehensive literature review, exploring existing research on graduate salaries, the role of

socioeconomic and demographic factors, and previous applications of machine learning in salary prediction. Section 3 outlines the research methodology, detailing the dataset, pre-processing steps, feature selection, and model evaluation criteria. Section 4 discusses the results and analysis, including the performance of different machine learning models, insights from SHAP feature importance analysis, and validation using statistical tests. Section 5 provides a critical discussion, addressing key findings, the implications of institutional prestige on salary outcomes, and the potential influence of AI on the future knowledge economy. Finally, Section 6 concludes this study by summarising key insights, identifying limitations, and suggesting directions for future research.

2. Literature Review

The landscape of higher education has transformed in recent years, becoming accessible to a wider population than ever before. However, the concept of student satisfaction remains elusive, and there is a pressing need to understand this phenomenon better. One crucial aspect of student satisfaction is the impact of socioeconomic and demographic factors on graduate salaries, as the dwindling rate of graduate salaries is an alarming trend. The literature review explores existing studies on graduate salaries and socioeconomic and demographic factors that influence graduate salaries, and data analytic techniques used in previous studies. By examining and synthesising the literature, this section aims to provide a solid foundation for the investigation of the impact of socioeconomic and demographic factors on graduate salaries in this study.

2.1. Factors Affecting Graduate Salaries

2.1.1. Employability

According to research, graduate employability is a critical factor affecting graduate salaries. Hogan et al. [21] define employability as one's ability to gain and maintain a job. Employability skills, which include knowledge, competencies, experience, personality, emotional intelligence, and career learning, are essential in determining a graduate's employability [22]. Rosenberg et al. [23] identified eight employability skills, including basic literacy, numeracy, critical thinking, management, leadership, interpersonal, information technology, systems thinking, and work ethic disposition. Similar other research [24] indicated that employers look for specific employability skills in graduates as these skills directly affect salaries. However, a limitation of current practices is the lack of a standardised measurement variable for quantifying employability skills, leading to inconsistencies in evaluating their impact on salaries [25].

Several prior studies have used quantitative econometric models to examine the relationship between employability skills and graduate salaries. Walker and Zhu [5] applied longitudinal regression models to track salary progression, demonstrating that graduates with industry-relevant skills experience faster wage growth. Similarly, Holmes and Mayhew [6] employed binomial logistic regression to assess how soft skills and career competencies influence early career earnings. These methodologies highlight the increasing role of skill-based hiring in determining salaries, emphasising the need for better employability skill assessments in graduate outcomes research.

2.1.2. Discipline and Gender

The ongoing debate about whether discipline and gender have a substantial impact on graduate salaries has been a topic of discussion for over a decade. Research has shown that multiple variables influence a graduate's salary and that merely possessing a degree is not enough. Light and Strayer [26] developed a taxonomy that identified 11 factors contributing to wage disparities among university graduates. Their findings revealed that

degree type, college of origin, and degree level (bachelor's, master's, and doctorate) all impact graduate salaries.

Zhang [27] found that the quality and discipline of a graduate's college education also play a role in determining their salary. Their research demonstrated that all disciplines, except life sciences and history, experienced a steady increase in pay over time among different genders. However, fields such as engineering, business management, and mathematics saw larger pay increases. Furthermore, Taniguchi [28] discovered that the negative effect on pay is more significant for men than for women, possibly because women tend to benefit less from their higher education as time passes. From a methodological standpoint, studies on discipline and gender wage gaps have relied on regression models and econometric decomposition methods. For example, Quadlin et al. [29] applied Blinder–Oaxaca decomposition models to quantify gender-based pay disparities across disciplines. Their findings suggest that while women earn as much as men in STEM fields during their first year, the pay gap widens over time. Additionally, nearly 100% of graduates in computer sciences, mathematics, engineering, and architecture earn above the national average salary, though employment rates among men remain 12% higher than among women [30]. These studies underscore the need for more nuanced wage gap models that integrate career progression trends and industry-specific factors.

In conclusion, it has been established in the literature that the discipline and gender are significant contributors to the graduate salary.

2.1.3. Socioeconomic Background

The influence of the socioeconomic background on the growth of a graduate's salary has been a topic of interest for researchers. Macmillan et al. [31] found that less advantaged young people are less likely to enter high-paying professional careers after graduation compared to their more advantaged peers, and this differential persists even when individual academic achievement levels are similar. Their study employed regression modelling and non-parametric analysis to assess how parental occupation and school type impact salary growth. Despite controlling for various background factors, they concluded that private school graduates have faster wage progression than those who attended public schools, suggesting that the educational background contributes to long-term salary disparities [8]. However, Duta et al. [7] challenged these findings, arguing that the socioeconomic status alone does not fully determine salary growth. Their research utilised decision tree algorithms and machine learning-based classification models to examine the impact of age, degree classification, and university type on wage trajectories. Their study found that age at graduation plays a more significant role than parental income in predicting early career earnings. These findings indicate that advanced machine learning techniques can provide richer insights than traditional regression models by capturing complex, non-linear interactions between socioeconomic and demographic factors.

Given the contradictions in existing research, there is a need for further empirical studies that integrate both econometric and machine learning methodologies to establish a clearer understanding of the major determinants of graduate salaries. Future research should consider hybrid modelling approaches, incorporating deep learning frameworks alongside traditional statistical techniques to improve wage prediction accuracy.

3. Methodology

3.1. Study Design, Data Collection, and Selection Criteria

This study adopts a quantitative, retrospective observational design, utilising secondary data from the Higher Education Statistical Agency (HESA) to analyse the impact of socioeconomic and demographic factors on graduate salaries. The dataset spans a three-

year period (2017/18–2019/20) and includes key variables such as the institution attended, parental education, employment basis, domicile, and salary classification. To ensure data quality and relevance, inclusion criteria were applied, restricting the analysis to graduates with recorded salary information and complete demographic and socioeconomic details. Conversely, data points with missing salary values, incomplete records or extreme salary outliers were excluded to prevent skewed results. This study employed machine learning models (e.g., decision trees, random forests) alongside statistical validation techniques (ANOVA, SHAP analysis) to uncover salary determinants and assess their significance.

In this work, the following steps were sequentially adopted for analysis.

3.2. Machine Learning Model Selection

Salary prediction has been analysed using different machine learning models. To identify the best candidate for the study needs, we compared multiple models by evaluating the performance of these models to determine the most suitable model for predicting graduate salaries.

3.2.1. Logistic Regression (LR)

LR analysis has increasingly been utilised as a statistical tool in analysing graduate salaries, particularly over the last two decades [32]. LR is widely recognised as the statistical method of choice when predicting binary outcomes, such as the likelihood of a graduate's salary being above or below a certain threshold, based on one or more independent variables. In this approach, the model is designed to predict the probability of each possible outcome category, while controlling for one or more independent variables. This method is used when the outcome variable is multiclass and is often referred to as multinomial or polychotomous logistic regression. The LR model is in the form of Equation (1) as it is expressed as a natural logarithm of the odds ratio:

$$In\left[\frac{P(Y)}{1-P(Y)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{1}$$

and

$$\left[\frac{P(Y)}{1-P(Y)}\right] = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k} \tag{2}$$

$$P(Y) = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k} - P(Y)e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k} \tag{3}$$

$$P(Y) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k}} \tag{4}$$

where P: probability function, $P(Y)$: the probability of the outcome Y, $In\left[\frac{P(Y)}{1-P(Y)}\right]$ is the log (odds) of the outcomes, Y is the multinomial output, $(X_1, X_2, \cdots, X_k)$ are the predictor variables and $(\beta_0, \beta_1, \beta_2, \cdots, \beta_k)$ are the regression model coefficient. LR, however, has certain limiting factors as it assumes linearity between the independent variables and logs odds of the dependent variable, requires independent observations, is sensitive to outliers, and is limited to binary/categorical outcomes. It requires a large sample size and assumes independence among the independent variables to avoid multicollinearity issues.

3.2.2. K-Nearest Neighbours (KNN)

The KNN algorithm is a popular and simple machine learning technique used for classification and regression tasks [33]. The basic idea behind KNN is to find the k-nearest data points to a given input data point and assign a label based on the majority class of those k-nearest neighbours. The distance metric used to calculate the distance between data

points can vary, but the Euclidean distance is a commonly used metric. First, the k-nearest neighbours: calculating the Euclidean distance between two points or tuples.

First data point: $X(1, X_{11}, X_{12}, \ldots, X_{1n})$

Second data point: $X(2, X_{21}, X_{22}, \ldots, X_{2n})$, using Equation (5)

$$d(X1, X2) = \sqrt{\sum_{i=1}^{n}(x_{i1} - x_{i2})^2} \quad (5)$$

However, it is important to note that KNN has some limitations. For instance, the choice of the k value can impact the accuracy of the model, and selecting the optimal k value may require experimentation [34]. Additionally, KNN may not perform well in cases where the dataset has a large number of variables or when the data are skewed [35]. In summary, while KNN can be a useful modelling technique for predicting the graduate salary, careful consideration of the data characteristics should be taken into consideration before selecting this approach. It is also important to assess the performance of the KNN model using appropriate evaluation metrics, such as the mean squared error or the coefficient of determination.

3.2.3. Linear Discriminant Analysis (LDA)

LDA is a statistical technique used for classification and dimensionality reduction [36]. LDA aims to find a linear combination of features that best separates the classes in a given dataset. The technique assumes that the features in the dataset are normally distributed and that the covariance matrices of each class are equal. To find the linear combination of features, LDA calculates the mean vectors and scatter matrices for each class. The scatter matrices represent the variability of the data within each class and are used to calculate the eigenvectors and eigenvalues of the data. The eigenvectors with the highest eigenvalues are then used to create a projection matrix that maps the original data onto a lower dimensional space. The resulting lower dimensional space can then be used for classification tasks. New data points are projected onto the same lower dimensional space, and their class membership is determined based on their proximity to the class centroids in this space [37]. LDA has been shown to perform well on a variety of classification tasks, particularly when the number of classes is small compared to the number of features. However, LDA assumes that the data are normally distributed and may not perform well if this assumption is violated.

3.2.4. Decision Tree

A decision tree is a supervised learning algorithm used for classification and regression tasks. A decision tree is a machine learning model that splits a dataset into smaller subsets based on feature conditions. Each feature used becomes a parent node and the data that it splits into are child nodes. This process is repeated until classification is reached. Decision trees are easy to understand as they can be visualised as flow charts, and they are also fast to train compared to other classifiers. It works by recursively partitioning the data into subsets based on the value of an attribute or feature to minimise the impurity of the subsets [38]. Each partition is represented by a binary decision made at a node of the tree. The resulting tree can be used for prediction by following the path from the root to a leaf node that corresponds to a specific set of conditions [39]. Decision trees have several advantages, including their interpretability, ability to handle both numerical and categorical data, and resistance to overfitting when properly pruned [40]. They are also efficient in training and can handle high-dimensional data.

One of the most popular decision tree algorithms is CART (Classification and Regression Trees) which builds a binary tree by recursively splitting the data into two subsets

based on the value of a single feature, such that the resulting subsets are as pure as possible concerning the target variable. The splitting criterion used by CART is typically the Gini impurity or the entropy, which measures the degree of impurity or randomness in a set of labels.

3.2.5. Random Forest (RF) Algorithm

The random forest is an ensemble method that combines multiple decision trees, each trained on a random subset of the data and a random subset of the features. The output of the random forest is the average or majority vote of the predictions made by the individual trees. Recent research has focused on improving the performance and interpretability of decision trees. For example, some studies have proposed new splitting criteria based on information theory, such as the Minimum Description Length (MDL) principle [41] or the Information Gain Ratio (IGR). Other studies have investigated ways to incorporate domain knowledge or expert rules into the decision tree learning process [30]. Still, others have explored methods to generate more compact and interpretable decision trees, such as rule-based pruning [42] or decision tree compression [39]. Furthermore, RF is an ensemble of machine learning models that are composed of many decision trees (without pruning) [43]. These trees are created by randomly sampling from the training data and using random subsets of features to determine the splits at each node. The final prediction of a random forest is made by aggregating the predictions of the individual decision trees.

The RF model is often less sensitive to changes in parameter settings compared to other predictors. The optimisation of an RF is usually based on two parameters: the number of variables used in splitting a node (mtry) and the number of trees in the model (ntrees). The optimal value of mtry is determined by testing all possible values. According to research by Breiman, the RF model's generalisation error decreases as the number of trees increases, a property that is not present in most other classifiers [44]. In other words, the model becomes more accurate as the value of ntrees increases. The optimisation of ntrees involves finding a balance between classification accuracy and computational efficiency.

3.2.6. Gaussian Naïve Bayes (GNB)

GNB is a classification algorithm that is based on the Bayes Theorem, it is one of the widely used algorithms in data mining, which states that the probability of an event occurring is equal to the prior probability of the event multiplied by the likelihood of the event given some observations. Naïve Bayes is a useful classifier that is used widely in many applications such as text categorisation [45] and data stream classification [46]. Bayesian classifier works based on the Bayesian rule and probability theorem. The Bayesian classifier operates under two assumptions. The first assumption is that, given the class label, the attributes are conditionally independent. The second assumption is that no latent attribute influences the prediction process for the label. A given vector $(x_1, \ldots, x_n)$ represents the n attributes of the instance x. Let c represent the class label of the instance x. The probability of x given C can be computed in Equation (6), where C is a class label:

$$p(C) = \prod_{i=1}^{n} p(x_1|C) \tag{6}$$

The conditional independence assumption of attributes in naïve Bayes is often incorrect for real-world problems, except in situations where the attributes are derived from independent processes. To address this, methods have been proposed to improve the assumption in naïve Bayes. In the case of GNB, the likelihood of the event is assumed to be a Gaussian distribution. This means that for each class, the classifier assumes that

the feature values are normally distributed in Equation (7) with a mean (μ) and standard deviation (σ).

$$p(x_1|C) = \frac{1}{\sqrt{2\pi\sigma_{c,i}^2}} e^{(-\frac{(x_i-\mu_{c,i})^2}{2\sigma_{c,i}^2})} \quad (7)$$

One of the limitations of the model is that it is sensitive to missing data, it has the possibility of struggling with data with high-dimension and the assumption of independent attributes is often unrealistic in real-world data. This can lead to poor performance of the model.

3.2.7. Gradient Boosting Algorithm

Gradient boosting is one of the most popular and effective machine learning algorithms for a wide range of tasks [47]. The basic idea behind gradient boosting is to minimise a loss function by adding weak learners in a greedy, forward stage-wise manner. At each stage, a new learner is fit to the residual errors of the previous learners, intending to reduce the error of the overall model. The gradient of the loss function concerning the output of the previous learners is used to guide the construction of the new learner. According to Chen and Guestrin [48], gradient boosting can be applied to both regression and classification problems and is particularly effective for tasks such as ranking, recommendation, and text classification. It has also been used in a variety of fields, including finance, biology, and astronomy. GBDT's drawback is its single-tree approach for enhancing models, rather than foresting.

3.2.8. Support Vector Machine (SVM) Algorithm

SVMs are a type of supervised learning algorithm that can be used for classification tasks. The goal of SVMs is to find the hyperplane in a high-dimensional space that separates different classes of data points. The hyperplane is defined by the support vectors, which are the data points closest to it, and is known as the separating hyperplane. They are particularly useful for handling high-dimensional spaces and dealing with non-linearly separable data using the kernel trick. A given separating hyperplane is defined based on the most significant minimum distance that a group of data frame observations has from that hyperplane [49]. A function, f(x), to define the sigmoid support vector classifier can be written as the following Equation (8):

$$f(x) = \frac{1}{1 + e^{-(\alpha + \sum_{i=1}^{n} \beta_i . \sum_{j=1}^{k} X_{ij} X_{i'j})}} \quad (8)$$

where

i, i': Represent indices of observation in the dataset;

X_{ij}: The j-th feature of the i-th sample;

$X_i{}'_j$: The j-th feature of a transformed or alternative representation of i-th sample;

$X_i{}'$: A modified representation of the feature vector X_i;

e: Represents the mathematical constant, approximately equal to 2.718.

There are n parameters, β_i being one for each observation $i = 1, \ldots, n$. The term $\sum_{j=1}^{k} X_{ij} X_{i'j}$ is the inner product between two observations i and $\hat{i}$, whose nomenclature can also be given by $\langle X_i, X_i' \rangle$. To estimate the parameters α and $\beta_i, \ldots, \beta_n$, the researcher has to define the $\frac{n(n-1)}{2}$ for inner products $\langle X_i, X_i' \rangle$ between all pairs of observations present in the data frame. $(n-1)/2$ represent the number of unique pairwise inner products $\langle X_i, X_i{}' \rangle$ that need to be computed when analysing feature interactions.

One of the limitations associated with SVMs is sensitivity to the choice of the kernel and hyperparameters, and finding the right combination can be difficult as they are not robust to noise and can overfit the data of the model [50].

3.2.9. Neural Networks

Deep learning is a subfield of ML that has been applied to various tasks, including image classification [51], segmentation [52], object detection [53], remote sensing [54], speech recognition, natural language processing [55], among others. It has been shown to achieve good performance in these tasks. A neural network is a type of ML algorithm that is inspired by the structure and function of the human brain. It is composed of a network of interconnected nodes, called artificial neurons, which receive input through synapses connected to axons; it then processes and transmits the information. In a neural network, signals passing through artificial neurons are mathematically represented as a combination of inputs and weights that are multiplied together to form a weighted sum. This sum is then passed through an activation function to generate an output. The activation function determines the output of the neuron based on the weighted sum and allows the network to model complex relationships between the inputs and the output (see Figure 1).

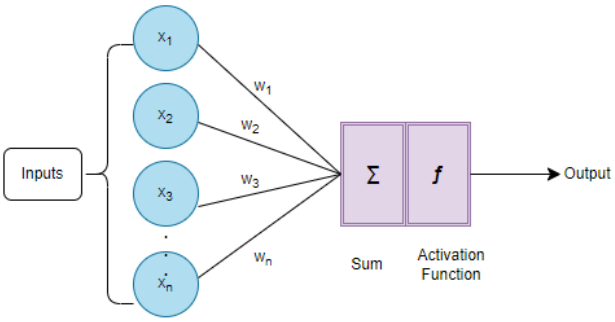

Figure 1. Mathematical representation of the neural network.

Neural networks have several limitations that can affect their performance. Neural networks can be sensitive to noise and outliers in the data, which can negatively impact their performance. This can be mitigated by pre-processing the data to remove noise and outliers or by using more robust algorithms. And it is prone to overfitting.

3.2.10. Adaptive Boosting Classifier Algorithm (AdaBoost)

The AdaBoost algorithm is an iterative approach that involves training multiple weak classifiers on the same dataset, which are then combined to form a stronger final classifier [56]. Throughout the process, the weights of the classifiers are adjusted to increase their accuracy and reduce their number for better specificity. The resulting model can then be used for classification. The Haar feature-based AdaBoost algorithm can be summarised as follows: The sample to be entered $(x_1, y_1), \ldots, (x_i, y_i)$, where x_i is the input i^{th} sample, and y_i represents the corresponding attribute value. The number of positive and negative samples are a, b, n = a + b. This algorithm is particularly effective in situations where there are complex interactions between features, and it can reduce the risk of overfitting [57]. However, it can be sensitive to noisy data and outliers, and it requires careful tuning of its parameters to achieve optimal performance.

3.3. Analysis of Variance (ANOVA)

ANOVA is a statistical technique used to test the hypothesis that the mean of a dependent variable is equal across different levels of an independent variable. It is used to determine whether there is a significant difference between the means of two or more

groups. ANOVA is based on the concept of variance, which is a measure of the dispersion of a set of data around its mean. ANOVA tests the null hypothesis that the means of the groups are equal against the alternative hypothesis that at least one of the means is different. There are several types of ANOVA, including one-way ANOVA, two-way ANOVA, and repeated measures ANOVA. The choice of ANOVA type depends on the characteristics of the data and the research question. The equation for the one-way model is in Equation (9):

$$y_{ij} = \mu_i + \epsilon_{ij}\{i = 1, 2, \ldots, pj = 1, 2, \ldots, n\} \quad (9)$$

where y_{ij} is the $ijth$ observation, μ_i is the mean of the i-th factor level or treatment, and ϵ_{ij} are a random error. and Table 1 shows the ANOVA table while Table 2 gives a summary of all the selected models and reason for inclusion.

Table 1. One-way ANOVA, where SSB is the sum of squares between groups, MSB is mean squares within groups, SSW is the sum of squares within groups, MSW is the mean square within groups, N is total number of observations and p is the number of groups.

Source of Variation	Sum of Squares	df	Mean Square	F
Between groups	SSB	(p-1)	$MSB = \frac{SSB}{(p-1)}$	$\frac{MSB}{MSW}$
Error (within groups)	SSW	(N-p)	$MSB = \frac{SSW}{N-p}$	
Total	SST	(N-1)	$s^2 = \frac{SST}{N-1}$	

Table 2. Graduate salary prediction model summary.

Model	Type	Key Characteristics	Strengths in Graduate Salary Prediction	Limitations	Reason for Selection in This Study
Logistic Regression (LR)	Classification	Assumes a linear relationship between independent variables (demographic/socioeconomic factors) and salary groups.	Simple and interpretable; useful for understanding relationships between predictors and salary categories.	Assumes linearity; may not capture complex relationships between factors like institutional prestige and salaries.	Used as a baseline model to compare performance with more complex ML methods.
K-Nearest Neighbours (KNN)	Classification	Classifies salaries based on the most similar historical cases; uses distance metrics.	Effective for detecting local salary patterns (e.g., institution-based salary clusters).	Computationally expensive for large datasets (1.87 M+ records); sensitive to irrelevant features.	Tested for its ability to capture salary variations based on demographic similarities.
Linear Discriminant Analysis (LDA)	Classification	Reduces dimensionality while maintaining class separability; assumes normal distribution.	Useful for analysing how multiple socioeconomic factors interact to classify salaries.	Assumes normality; may not perform well with non-Gaussian salary distributions.	Applied for dimensionality reduction and exploratory analysis of salary classifications.
Decision Tree (DT)	Classification	Recursive partitioning method; identifies key decision rules (e.g., "Did graduate attend a top-tier university?").	Highly interpretable; detects hierarchical salary determinants (institution > job qualification > parental education).	Prone to overfitting without pruning; splits may be biased towards dominant features.	Identified as the best-performing model due to its ability to classify salary groups with high accuracy.
Random Forest (RF)	Classification	Ensemble of decision trees; aggregates multiple models to improve accuracy and reduce overfitting.	Handles large datasets effectively; identifies the most influential salary determinants across multiple trees.	Less interpretable than a single decision tree; computationally intensive.	Provided the highest accuracy and robust feature importance ranking (validated by SHAP analysis).
Gaussian Naïve Bayes (GNB)	Classification	Probabilistic model assuming feature independence; calculates salary probabilities per demographic group.	Works well on high-dimensional categorical data (e.g., ethnicity, institution type).	Assumes feature independence, which may not hold (e.g., parental education and socioeconomic status are correlated).	Included as a benchmark for probabilistic classification of graduate salary categories.

Table 2. *Cont.*

Model	Type	Key Characteristics	Strengths in Graduate Salary Prediction	Limitations	Reason for Selection in This Study
Gradient Boosting (GB)	Classification	Boosting technique that sequentially improves salary classification by correcting previous errors.	Strong predictive performance; handles interactions between features like "field of study × institution reputation".	Prone to overfitting; requires extensive tuning.	Applied to test boosting-based performance improvement techniques.
Support Vector Machine (SVM)	Classification	Maximises margin between salary classes; effective in high-dimensional spaces.	Can model non-linear salary patterns using kernel trick (e.g., interaction of job qualification and domicile).	Computationally expensive on large datasets; sensitive to hyperparameter tuning.	Evaluated for its ability to classify salaries in complex feature spaces.
Neural Network (NN)	Classification	Learns hierarchical patterns in salary determinants; captures non-linear relationships.	Can detect deep salary trends (e.g., how combinations of ethnicity, institution, and job type influence pay).	Requires extensive data; hard to interpret; high computational cost.	Included to compare deep learning techniques with traditional ML models.
AdaBoost (ADA)	Classification	Assigns more weight to misclassified salary instances; iteratively improves model accuracy.	Improves weak learners; enhances salary predictions for minority groups.	Sensitive to noisy data; requires many iterations.	Used to assess the performance of boosting-based classifiers in handling salary disparities.
ANOVA (Analysis of Variance)	Statistical Analysis	Tests whether salary differences across demographic groups (e.g., gender, ethnicity) are statistically significant.	Validates machine learning findings; confirms whether socioeconomic factors significantly impact salaries.	Assumes homogeneity of variance; outliers can impact results.	Used to confirm statistical significance of identified salary determinants before ML modeling.

3.4. The Data Science Pipeline

Analysing the influence of socioeconomic and demographic factors on graduate salaries demands a systematic data science approach. Leveraging HESA data, this process starts with comprehensive data collection and Exploratory Data Analysis (EDA) to uncover underlying patterns. Subsequent data preparation and cleaning ensure integrity, followed by feature engineering to enhance insights. Model training employs machine learning algorithms, iteratively refined through validation and hyperparameter tuning for optimal performance. Interpretation unveils model workings using techniques like SHAP, while feature selection distils crucial variables. This journey transforms raw data into actionable insights, empowering decision-makers as presented in Figure 2. Through this approach, the intricate impact of socioeconomic and demographic factors on graduate salaries is revealed, aiding informed decision-making in academia and beyond.

Explanation of the Machine Learning Pipeline (Figure 2)

The machine learning pipeline in this study follows a structured approach to analysing the impact of socioeconomic and demographic factors on graduate salaries. The first stage involves data collection and pre-processing (Table 3), where secondary data from the Higher Education Statistical Agency (HESA: 2017/18–2019/20, https://www.hesa.ac.uk/data-and-analysis/students, accessed on 4 February 2025) are utilised. This stage includes handling missing data through imputation or removal, encoding categorical variables such as gender and institution type, and applying normalisation and outlier detection to maintain data integrity.

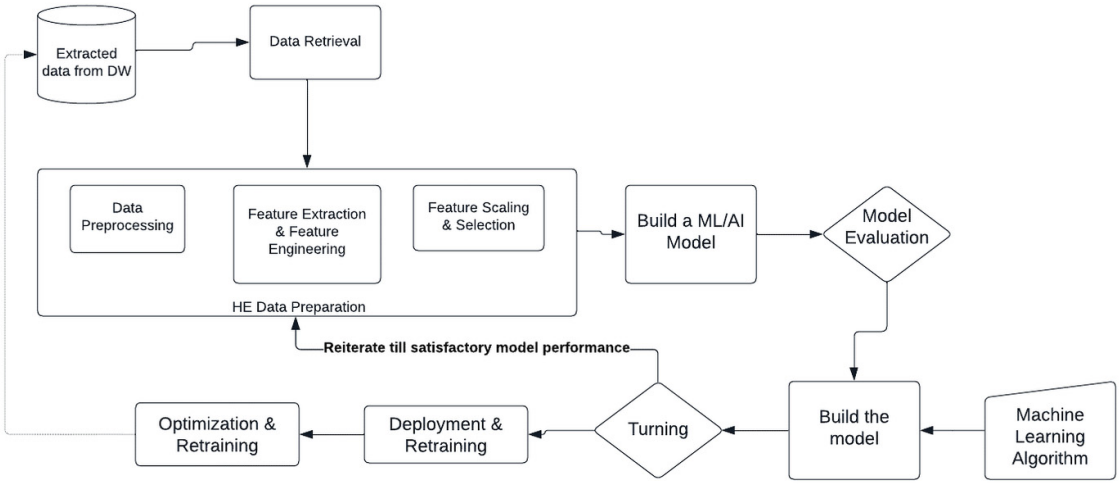

Figure 2. Machine Learning Pipeline.

Table 3. Pre-processed categorical variables and their encodings.

Category	Specific Variables	Pre-Processed
Personal Characteristics	Sex	Male = 1, Female = 0, Others = 2
	Disability	No known disability = 1, Known disability = 0
Socioeconomic	Socio economic classification	Higher managerial and professional occupations = 0, Not classified = 5, Intermediate occupations = 1, Lower managerial and professional occupations = 2, Routine occupations = 6, Lower supervisory and technical occupations = 3, Semi-routine occupations = 7, Small employers and own account workers = 8, Never worked and long-term unemployed = 4
	Parental education	Yes = 4, No = 2, Information refused = 1, No response given = 3, Do not know = 0
Demographic	Domicile	England = 0, Non-European Union = 1, Other European Union = 4, Scotland = 6, Wales = 7, Northern Ireland = 2, Other UK = 5, Not known = 3
	Ethnicity	White = 6, non-UK = 3, Asian = 0, Black = 1, Mixed = 2, Other = 5, Not known = 4
Academic level/Qualification	Level of study	First degree = 0, Postgraduate (taught) = 1, Other undergraduate = 2, Postgraduate (research) = 3
	Class of first-degree	Classification not applicable = 0, Upper second-class honours = 6, First class honours = 2, Lower second class honours = 3, Unclassified third class honours/Pass = 4, FE level qualification = 1
	Mode of study	Full time = 0, Part time = 1
	Academic Year	2017/18 = 0, 2018/19 = 1, 2019/20 = 2
	Interim study	No significant interim study = 0, Unknown interim study = 2, Significant interim study = 1
Employment type/location	Location of work	England = 0, Northern Ireland = 1, Overseas = 4, Wales = 6, Scotland = 5, Not known = 2, Other UK = 6
	Employment basis	On a permanent/open-ended contract = 3, On a fixed-term contract lasting 12 months or longer = 1, On a fixed-term contract lasting less than 12 months = 2, On a zero-hour contract = 4, On an internship = 5, Temping (including supply teaching) = 7, Other = 6, Volunteering = 8, Not known = 0
	Employment mode	Full time = 0, Part time = 1
	Most important activity	Paid work for an employer = 4, Engaged in a course of study, training or research = 3, Running my own business = 6, Self-employment/ freelancing = 7, Developing a creative, artistic, or professional portfolio = 1, Unemployed and looking for work = 9, Retired = 5, Doing something else = 2, Caring for someone (unpaid) = 0, Voluntary/unpaid work for an employer = 10, Taking time out to travel—this does not include short-term holidays = 8
	Publication main activity	Full-time employment = 1, Part-time employment = 4, Employment and further study = 0, Full-time further study = 2, Part-time further study = 5, Unemployed = 6, Other including travel, caring for someone, or retired = 3, Voluntary or unpaid work = 11, Unemployed and due to start work = 8, Unknown pattern of employment = 9, Unemployed and due to start further study = 7, Unknown pattern of further study = 10
	Salary groups	Minimum wage = 2, Living wage = 0, Median wage = 1, Top wage = 3
Markers	Subject type marker	Non-science = 1, Science = 0
	State school marker	State-funded school or college = 1, Unknown or not applicable school type = 2, Privately funded school = 0

Following pre-processing, Exploratory Data Analysis (EDA) is conducted to identify trends in salary distribution across demographic factors. EDA techniques help detect potential biases or anomalies in the dataset and visualise salary group classifications, such as minimum wage, living wage, median wage, and top wage.

The next stage, model training and evaluation, involves comparing ten machine learning models, including decision trees, random forests, logistic regression, and neural networks. Due to class imbalance in salary groups, oversampling and undersampling techniques are employed to improve predictive performance. The models are evaluated based on accuracy, precision, recall, and F1-score, with the decision tree model demonstrating the best performance in terms of accuracy and interpretability. To ensure reliability, this study incorporates interpretability and validation techniques, including SHAP analysis to explain feature contributions, ANOVA to validate the statistical significance of key socioeconomic factors, and expert validation from industry professionals to confirm practical relevance.

The final stage, policy recommendations and insights, translates the research findings into actionable strategies for universities, policymakers, and employers. These recommendations highlight interventions to address wage disparities based on socioeconomic background and inform decisions that promote equitable career opportunities for graduates.

While this pipeline offers several strengths, including data-driven insights, robust statistical validation, and expert feedback, it also has limitations. Machine learning models provide complex, automated predictions, but some, like neural networks, are difficult to interpret. The large HESA dataset (1.87 M+ records) strengthens statistical power but introduces challenges such as potential biases, missing values, and inconsistencies in definitions. Feature selection using SHAP enhances transparency but is dependent on the completeness of the available data. Addressing class imbalances through resampling improves model performance, though oversampling may introduce synthetic noise, and undersampling can result in data loss. Lastly, ANOVA and expert validation ensure both statistical and real-world credibility, but expert opinions remain subjective and may not fully account for statistical nuances.

Enhancing the discussion around Figure 2 (machine learning pipeline) strengthens the clarity, methodological rigour, and overall impact of this research. This expanded explanation ensures that readers, reviewers, and policymakers fully understand the analytical process, its strengths, and its limitations, allowing for informed decision-making in higher education and labour market policies.

3.4.1. Data Overview

The data used for this research work are survey data conducted, monitored, and collected by HESA and ranges from 2017/18 to 2019/20. The data include approximately 28 variables and 1,871,245 data points, covering various factors that may influence graduate salaries, such as socioeconomic and demographic variables.

3.4.2. Data Types and Outcome Variables

The outcome variable in this study is the graduate salary, which is categorised into four distinct groups: minimum wage, living wage, median wage, and top wage. These categories represent different salary ranges that graduates fall into, based on their earnings after completing their education. While the salary itself is a continuous variable, for this study, it has been discretised into these four groups, turning the problem into a multi-class classification task. The four salary categories allow for a more nuanced analysis of the factors influencing salary distribution. Instead of predicting a single continuous value, the machine learning algorithms aim to classify graduates into one of the four predefined salary groups.

3.4.3. Limitations and Constraints of Secondary Data

Secondary data refer to data that have been collected, processed, and made available by an external organisation or researcher than being gathered firsthand for a specific study. In this research, the dataset is sourced from the Higher Education Statistical Agency (HESA), covering graduate salary records from 2017/18 to 2019/20. The use of secondary data provides a cost-effective and large-scale means of analysing socioeconomic and demographic influences on graduate salaries. However, it also introduces several limitations and constraints that can impact this study's reliability and interpretability. This study's reliance on secondary data introduces limitations like potential biases, missing information, and limited control over variable definitions. Enhancing secondary data with primary collection or advanced imputation techniques can mitigate these issues [58]. Additionally, cross-validation with external datasets can ensure generalisability.

3.4.4. Missing Data and Outliers

The presence of missing data and outliers was carefully addressed to ensure the robustness of the analysis. Rows containing blank or NA values were entirely dropped. The number of data removed during this process was minimal and did not significantly impact the final results. Outliers, a common issue in survey data, were also examined. In this study, some respondents provided unrealistically low or excessively high salary figures, potentially skewing the results. To address this, data transformation techniques were applied, ensuring that no salary group among the four predefined categories held an unfair advantage during the modelling process. This approach minimises the distortion caused by outliers while preserving the integrity of the data.

4. Results Analysis

This study employed ML/AI to examine how socioeconomic and demographic factors impact graduate salaries in the UK. Using survey data from HESA, this research focused on both qualitative and quantitative measures. The data underwent rigorous processing, including feature engineering for salary groups, and were used to train 10 machine learning classifiers. This study evaluated model performance, considering accuracy, precision, F1-score, and recall, while also assessing variable contributions using SHAP and examining interactions with ANOVA. The aim was to provide guidance for graduates and inform government policy decisions.

4.1. Model Training Results

The result presented in Table 4 highlights the performance metrics of different machine learning models applied to a graduate salary dataset. The models considered are LR (logistic regression), KNN (k-nearest neighbours), LDA (linear discriminant analysis), DT (decision tree), GNB (Gaussian naïve Bayes), GB (gradient boosting), SVM (support vector machine), NN (neural network), ADA (AdaBoost), and RF (random forest). The metrics used to evaluate the models' performance are precision, F1-score, recall, and accuracy.

In this case, Random Forest has achieved the highest accuracy of (0.72) as seen in Table 4; however, accuracy alone does not necessarily indicate the best-performing model, as it may be biased towards the majority class. Given the imbalance in the dataset, a more comprehensive evaluation is required. This has been analysed using a confusion matrix shown in Figure 3, which highlights the distribution of prediction across four salary categories: minimum wage = 0, living wage = 1, median wage = 2, and top wage = 3. The confusion matrix provides deeper performance by showing how many instances of each salary group were correctly classified and where misclassification occurred.

Table 4. Preliminary training results.

Original Classification Data				
	Model Evaluation Index			
Model	**Precision**	**F1-Score**	**Recall**	**Accuracy**
Logistic Regression	0.49	0.45	0.46	0.54
K-Nearest Neighbours	0.62	0.61	0.61	0.66
Linear Discriminant Analysis	0.49	0.46	0.46	0.54
Decision Tree	0.66	0.61	0.62	0.69
Gaussian Naïve Bayes	0.48	0.47	0.47	0.52
Gradient Boosting	0.64	0.59	0.59	0.67
Support Vector Machine	0.47	0.45	0.45	0.54
Neural Network	0.09	0.13	0.25	0.37
AdaBoost	0.56	0.51	0.52	0.6
Random Forest	0.68	0.67	0.67	0.72

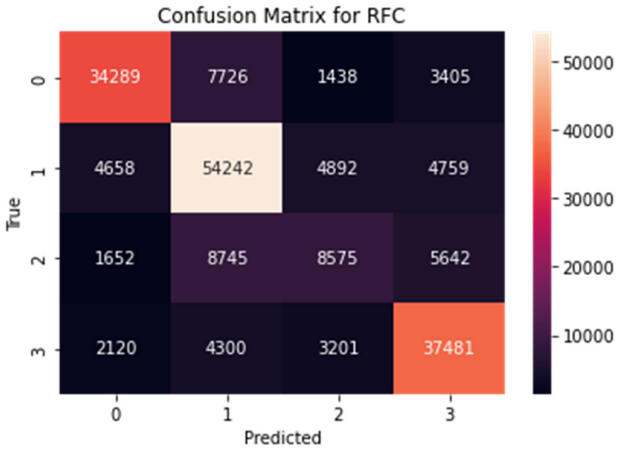

Figure 3. Confusion matrix salary classification using the random forest classifier.

A weighted selection approach is applied to determine the most suitable model, considering multiple performance indicators. Accuracy (30%) measures overall correctness but does not account for class imbalances, making it insufficient as a standalone criterion. F1-Score (30%) balances Precision and Recall, ensuring that both false positives and false negatives are minimised. Recall (20%) is crucial for correctly identifying all salary categories, especially under-represented ones, while Precision (20%) enhances the reliability of positive classifications, reducing incorrect high-salary predictions. Based on this approach, Random Forest emerges as the best-performing model, achieving the highest overall F1-Score (0.67), maintaining strong Precision and Recall, and offering high interpretability through feature importance analysis.

The matrix shows the number of predictions for each true class that were correctly classified (on the diagonal) and the number of predictions that were incorrectly classified (off the diagonal). Overall, the model has relatively high accuracy for class 0 and class 3, as indicated by the high numbers on the diagonal for these classes. However, the model has lower accuracy for class 1 and class 2, as indicated by the relatively high numbers of the diagonal for these classes. The model's performance is impacted by relatively high numbers of incorrect predictions between classes 1 and 3, leading to challenges in distinguishing between them and ultimately lowering the precision and accuracy of the model.

Resampling is performed to solve the sample imbalance and incorrect prediction between the true classes. The sampling method is divided into undersampling, oversampling, and combined sampling. For the sake of this work, we shall be interested in only undersampling and oversampling. Undersampling and oversampling techniques play pivotal roles in mitigating the challenges posed by imbalanced datasets, where one class

significantly outnumbers the others. These methods are not only of interest but are essential strategies for improving model performance and addressing biases inherent in imbalanced data distributions. Undersampling involves randomly removing instances from the majority class to match the size of the minority class, thereby reducing the dominance of the majority class. This process ensures a more balanced representation of classes in the dataset, facilitating better model learning and discrimination between classes. On the other hand, oversampling techniques increase the representation of the minority class by replicating or synthesising instances. By amplifying the presence of the minority class, oversampling prevents the model from being overshadowed by the majority class and improves its ability to generalise to unseen instances [59].

Furthermore, undersampling and oversampling help in enhancing model generalisation by preventing overfitting and underfitting. Undersampling prevents the model from being overly biased towards the majority class, leading to better generalisation performance on unseen data [60]. Similarly, oversampling ensures that the model adequately learns the characteristics of the minority class, thereby improving its ability to generalise across both classes. Moreover, these techniques also offer computational advantages. Undersampling reduces the computational burden by decreasing the dataset size, while oversampling, although increasing the dataset size, can still be computationally efficient compared to alternative methods of expanding the minority class representation [58].

By addressing issues related to data skewness, enhancing model generalisation, and mitigating computational costs, undersampling and oversampling techniques ensure fair, reliable, and effective model performance across various applications. Thus, their significance extends beyond mere interest, underscoring their essential role in the realm of imbalanced data analysis.

Table 5 provides an improvement model training result for the ten classifiers used to predict the impact of the various considered variables on the graduate salary in the UK. For both sampling techniques used, the evaluation metrics used are precision, recall, F1-score, and accuracy. For oversampling, the highest performing models based on F1-score are Decision Tree, Random Forest, and Gradient Boosting, with F1-scores of 0.85, 0.67, and 0.62, respectively. Decision Tree has the highest precision and recall scores, indicating it has correctly classified all positive cases. Decision Tree has the highest accuracy of 0.85.

For undersampling, the highest performing models based on F1-scores are Decision Tree, Random Forest, and K-Nearest Neighbours, with F1-scores of 1, 0.67, and 0.61, respectively. Decision Tree has the highest precision, recall, and accuracy scores, indicating it has correctly classified all positive cases. Comparing the results of the two-sampling technique, the models trained on undersampled data outperform the models trained on oversampled data or the preliminary training result, with Decision Tree and Random Forest performing best in both cases. This suggests that undersampling the majority class could be a better approach to this problem.

Overall, the evaluation tables suggest that Decision Tree and Random Forest are the best models for predicting the impact of socioeconomic and demographic factors on the graduate salary in the UK, with Decision Tree performing slightly better than the Random Forest in both oversampled and undersampled datasets, Figure 4 collaborate the study findings.

Table 5. Improved model training result.

	Model	Model Evaluation Index Precision	Recall	F1-Score	Accuracy
Oversampling	Logistic Regression	0.5	0.5	0.49	0.5
	K-Nearest Neighbours	0.61	0.61	0.6	0.62
	Linear Discriminant Ana.	0.51	0.5	0.48	0.5
	Decision Tree	0.86	0.85	0.85	0.85
	Gaussian Naïve Bayes	0.48	0.48	0.46	0.47
	Gradient Boosting	0.64	0.63	0.62	0.63
	Support Vector Machine	0.64	0.65	0.64	0.65
	Neural Network	0.25	0.27	0.18	0.39
	AdaBoost	0.54	0.55	0.54	0.56
	Random Forest	0.67	0.67	0.67	0.71
Undersampling	Logistic Regression	0.5	0.5	0.49	0.5
	K-Nearest Neighbours	0.61	0.62	0.61	0.63
	Linear Discriminant.	0.52	0.5	0.47	0.47
	Decision Tree	1	1	1	0.99
	Gaussian Naïve Bayes	0.48	0.47	0.46	0.47
	Gradient Boosting	0.65	0.63	0.6	0.6
	Support Vector Machine	0.63	0.63	0.65	0.64
	Neural Network	0.27	0.3	0.23	0.42
	AdaBoost	0.54	0.55	0.54	0.56
	Random Forest	0.68	0.67	0.67	0.72

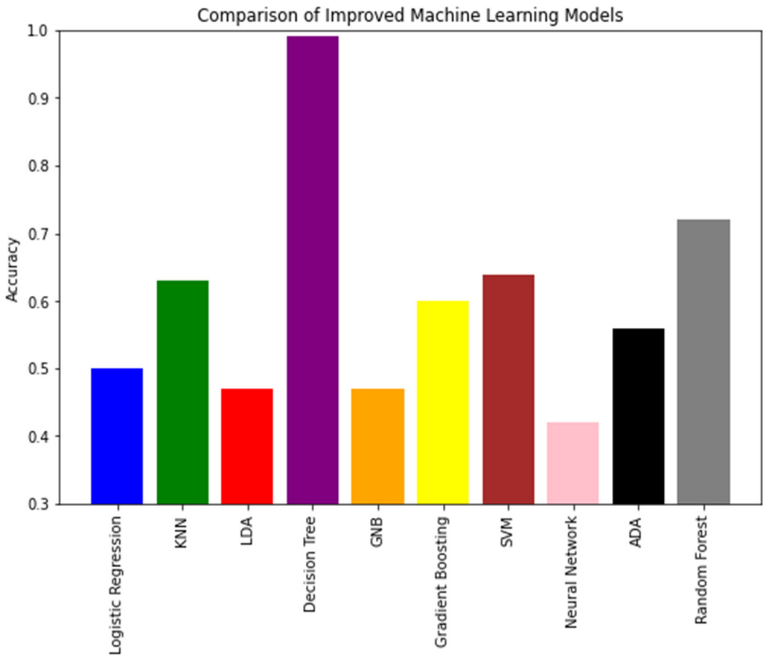

Figure 4. Compared to improved machine learning models.

The confusion matrix presented in Figure 5 highlights the performance of a classification model that has made predictions for four classes (0, 1, 2, and 3) on the graduate salary dataset using a random undersampling technique. Specifically, the confusion matrix can be interpreted as follows:

Class 0: There were 99,345 true instances of Class 0, and the model correctly predicted Class 0 for 99,345 of these instances. However, the model incorrectly predicted Class 1 for nine instances of Class 0, predicted Class 2 for seven instances of Class 0, and predicted Class 3 for 10 instances of Class 0.

Class 1: There were 99,347 true instances of Class 1, and the model correctly predicted Class 1 for 99,086 of these instances. However, the model incorrectly predicted Class 0 for 194 instances of Class 1, predicted Class 2 for 59 instances of Class 1, and predicted Class 3 for eight instances of Class 1.

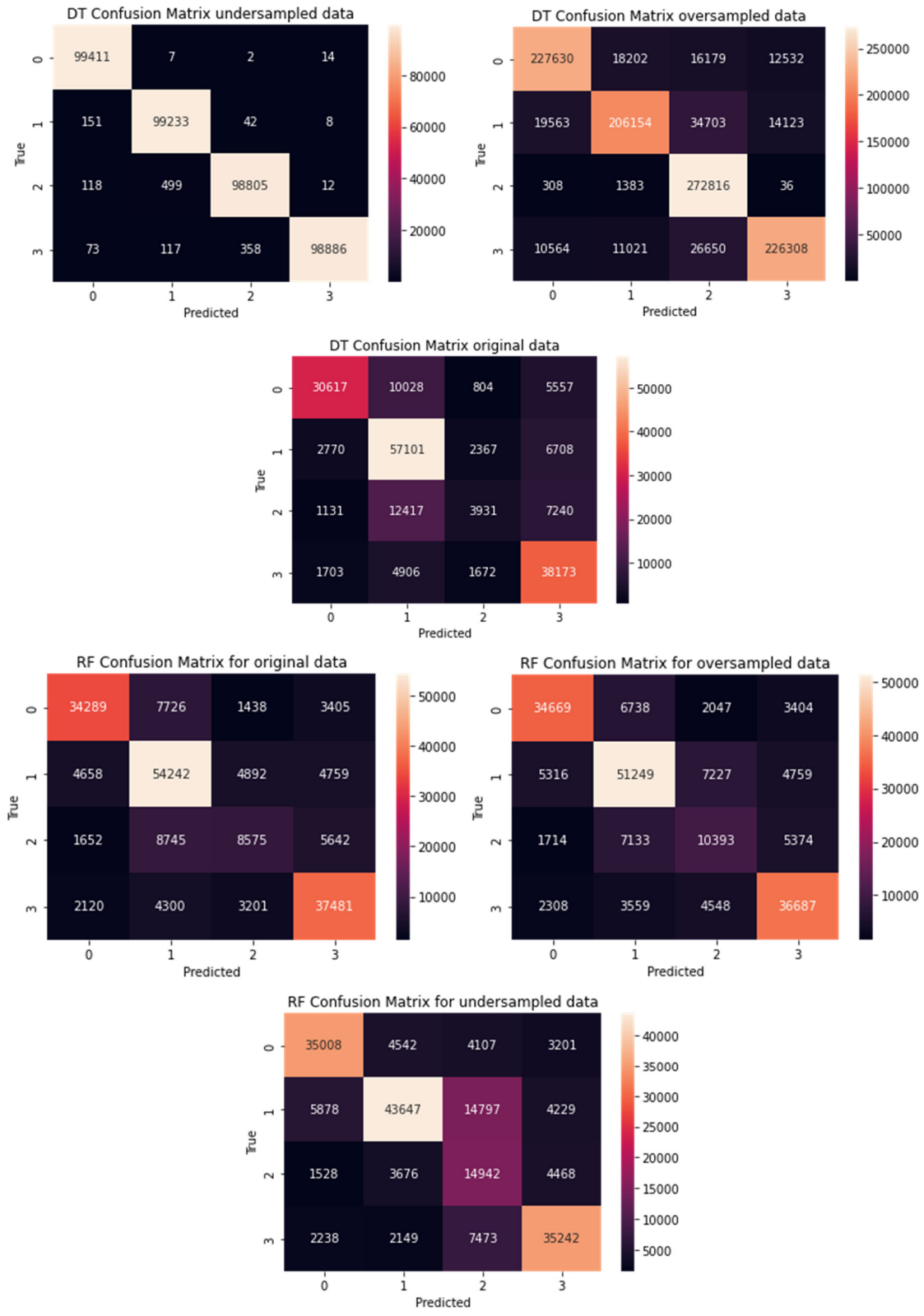

Figure 5. Confusion matrix salary classification using the random forest and decision tree classifiers.

Class 2: There were 99,416 true instances of Class 2, and the model correctly predicted Class 2 for 98,806 of these instances. However, the model incorrectly predicted Class 0 for 112 instances of Class 2, predicted Class 1 for 474 instances of Class 2, and predicted Class 3 for 24 instances of Class 2.

Class 3: There were 99,155 true instances of Class 3, and the model correctly predicted Class 3 for 98,628 of these instances. However, the model incorrectly predicted Class 0 for 71 instances of Class 3, predicted Class 1 for 113 instances of Class 3, and predicted Class 2 for 343 instances of Class 3.

Overall, the model has high accuracy for all four classes, as indicated by the high numbers on the diagonal for these classes. However, the model seems to have the most difficulty distinguishing between Classes 1 and 2, as there are relatively high numbers of incorrect predictions between these two classes. This can be seen in the confusion matrix, where the number of misclassified instances between Classes 1 and 2 is higher compared to other classes. The weighted average precision, recall, and F1-score for all classes are 1.00, indicating that Decision Tree has performed well.

4.2. Feature Importance Analysis Using Shapley Additive Explanations (SHAP)

In this next stage, feature importance has been analysed using the SHAP method (Table 6), which has been widely acknowledged for its consistency and interpretability in machine learning research [61]. SHAP provides an intuitive approach to understanding model predictions by attributing each feature's contribution to the outcome. These values indicate the relative contribution of each feature to the overall prediction made by the model. In the context of graduate salary data in the UK, this information can be useful in understanding which factors have the most significant impact on graduate salaries.

Table 6. Variable Importance.

Features	Importance
Provider	0.1450 (14.50%)
Age of qualifier	0.0941 (9.41%)
Academic year	0.0805 (8.05%)
Most important activity	0.0765 (7.65%)
Interim study	0.0677 (6.77%)
Employment basis	0.0665 (6.65%)
Socio economic classification	0.0634 (6.34%)
Domicile	0.0596 (5.96%)
Location of work	0.0437 (4.37%)
Publication main activity	0.0409 (4.09%)
Parental education	0.0363 (3.63%)
Qualification required for the job	0.0362 (3.62%)
Class of first degree	0.0343 (3.43%)
Ethnicity	0.0319 (3.19%)
Subject type marker	0.0214 (2.14%)
Sex	0.0204 (2.04%)
Level of study	0.0180 (1.80%)
Employment mode	0.0174 (1.74%)
State school marker	0.0170 (1.70%)
Disability	0.0162 (1.62%)
Mode of study	0.0099 (0.99%)

In the context of graduate salary data in the UK, feature importance appeared in the order listed in Table 6 where the "Provider" feature (institution attended) emerged as the most influential predictor, accounting for 14.50% of the model's predictive power. The second most important feature, "Age of qualifier", mirrors findings from studies on workforce demographics, which suggest that age impacts earning potential through experience and networking advantages. Furthermore, socioeconomic classification and parental education, though less impactful, contribute significantly to the prediction model. Other significant features include "Academic year", "Most important activity", "Interim study", and "Employment basis", all of which have importance values above 6%. These

factors could also impact a graduate's salary by influencing their skills, experience, and qualifications (see Figure 6 for the SHAP summary plot).

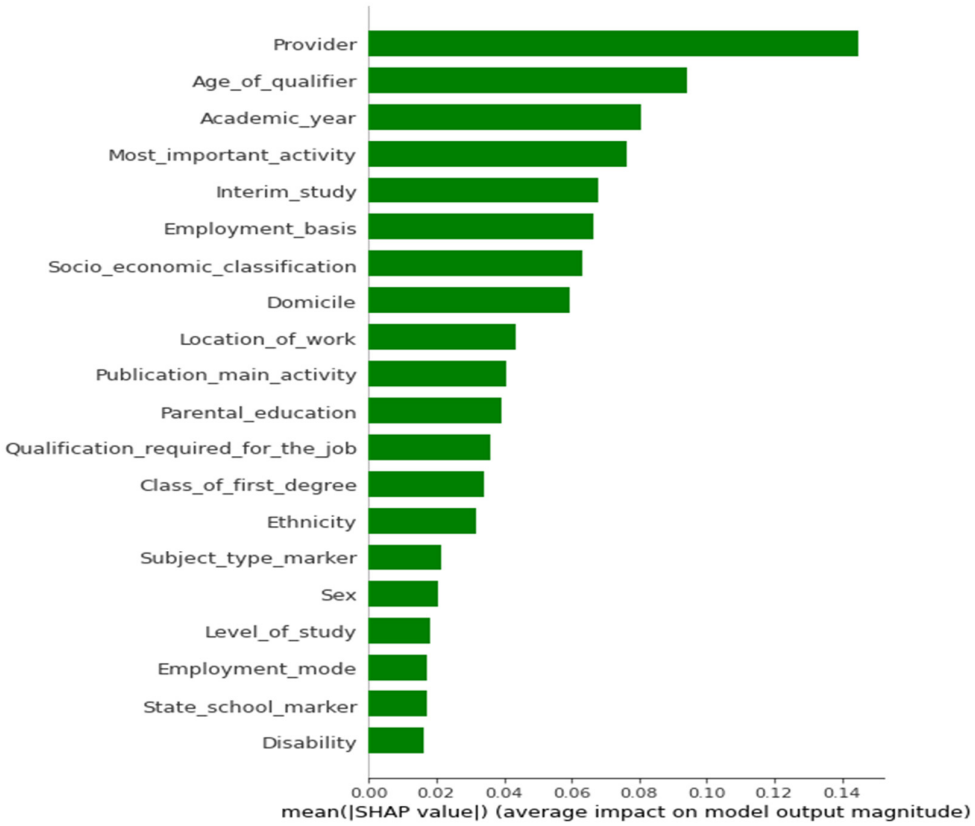

Figure 6. SHAP summary plot.

Overall, this analysis highlights the complexity of graduate salary data in the UK and the importance of considering multiple factors when predicting or explaining salary outcomes. While the university attended is a critical factor, other characteristics such as age, employment status, and academic performance also have a significant impact. Parental education, qualification required for the job, ethnicity, and the class of first degree are less important as revealed by the survey data but still contribute significantly to the model's prediction. Understanding these factors and investigating possible interactions and their relative importance can help graduates and employers make more informed decisions about education, training, and job opportunities.

4.3. Interaction of Significant Factors: An Analysis of Variance (ANOVA) Approach

Although the variable importance measured by the SHAP model is informative, it does not provide conclusive evidence about the statistical significance of the identified factors on graduate salaries. To determine whether these factors have a significant impact on graduate salaries and whether there are any interaction effects between them, we will use ANOVA with a significance level (α) of 0.05.

From Table 7, all independent variables (factors) imputed in the analysis of the variance model include Provider, Academic_year, Age_of_qualifier, Socio_economic_classification, Most_important_activity, Location_of_work, Employment_basis, Publication_main_activity, Domicile, Interim_study, Qualification_required_for_the_job, and Parental_education.

Table 7. The ANOVA table.

	sum_sq	df	F	PR (>F)
Provider	5.78×10^{10}	1	30.095345	4.11×10^{-8}
Academic_year	1.43×10^{13}	1	7427.0483	0.00×10^{0}
Age_of_qualifier	9.65×10^{12}	1	5023.6164	0.00×10^{0}
Socio_economic_classification	4.81×10^{11}	1	250.31424	2.24×10^{-56}
Most_important_activity	5.85×10^{12}	1	3048.1864	0.00×10^{0}
Location_of_work	2.71×10^{13}	1	14,116.678	0.00×10^{0}
Employment_basis	6.06×10^{11}	1	315.54089	1.37×10^{-70}
Publication_main_activity	3.00×10^{13}	1	15,602.032	0.00×10^{0}
Domicile	9.05×10^{13}	1	47,135.846	0.00×10^{0}
Interim_study	3.94×10^{13}	1	20,514.903	0.00×10^{0}
Qualification_required_for_the_job	6.10×10^{12}	1	3173.9941	0.00×10^{0}
Parental_education	1.55×10^{12}	1	809.56141	4.91×10^{-178}

Based on the ANOVA table, all of the variables have a p-value less than 0.05, indicating that they have a statistically significant impact on the graduate salary. This suggests that all of these variables should be considered when examining the impact on the graduate salary. It is also worth noting that some variables, such as Domicile and Publication_main_activity, have much larger F-statistics than others, indicating that they may have a greater impact on the graduate salary than other variables in the model. The results of the analysis indicate that most of the interactions are significant at the $p < 0.05$ level.

5. Discussion

This study aims to predict the graduate salary and identify the major factors that impact it using ten different machine learning methods. The importance of higher education is subject to an ongoing debate, with concerns about its ability to generate social mobility despite increasing investments by both the government and graduates. As a result, there is a growing demand for educational policies and guidance for prospective graduates to navigate the education system effectively.

This study employed survey data from HESA, consisting of 27 independent variables and one response variable (Salary). Ten machine learning models: logistic regression (LR), k-nearest neighbours (KNN), linear discriminant analysis (LDA), decision tree (DT), Gaussian naïve Bayes (GNB), gradient boosting (GB), support vector machine (SVM), neural network (NN), adaptive boosting (ADA), and random forest (RF) are evaluated for salary prediction. The successful implementation of the ten different models on the graduate salary data justifies the assumption that machine learning is suitable for modelling graduate salaries. After running the initial model using the original data, which was unbalanced, it was observed that the random forest classifier rendered the highest accuracy of about 72%, with a lot of false positive predictions about the graduate salary. It was clear that overfitting of the data on the model was visible hence requiring model improvement. To improve the precision, accuracy, and F1-score of the study model, we relied on a sampling method that comprised oversampling and undersampling.

In the oversampled data, it was observed that Decision Tree performed the best with an accuracy score of 85%. Random Forest also performed well with an accuracy score of 71%. Support Vector Machine and K-Nearest Neighbours had moderate accuracy scores of 65% and 62%, respectively. However, Gaussian Naïve Bayes had a relatively low accuracy score of 47%, indicating its poor performance on this particular dataset. Neural Network had the lowest accuracy score of 39%, indicating a need for further optimisation or tuning to improve its performance. In the undersampled data, Decision Tree also performed the best with a very high accuracy score of 99%. Random Forest also performed well with an

accuracy score of 72%. Support Vector Machine and K-Nearest Neighbours had moderate accuracy scores of 64% and 63%, respectively. Linear Discriminate Analysis and Gaussian Naïve Bayes, along with Neural Network, had relatively low accuracy scores.

By comparing the model evaluation with the original data, oversampled data, and undersampled data, we observed that the undersampled data provided the best improvement in the study model, leading to accurate predictions of graduate salaries as seen in the confusion matrix in Figure 6. Based on the findings, we can categorically state that the decision tree algorithm is the best machine learning algorithm for predicting graduate salaries in the UK, as it achieved the highest precision, F1-score, recall, and accuracy scores.

5.1. The Variable Impacts

Having established the decision tree to be this study's best-performing model, the input variables' (listed in Table 8) impact is further investigated to analyse how these variables impacted the model performance and also contributed to predicting the graduate salary. The SHAP variable importance technique was employed to achieve these.

Table 8. Input variables.

Input Variables
Academic_year
Publication_main_activity
Most_important_activity
Employment_mode
Interim_study
Sex
Domicile
Parental_education
State_school_marker
Socio_economic_classification
Age_of_qualifier
Disability
Ethnicity
Subject_type_marker
Level_of_study
Class_of_first_degree
Provider
Mode_of_study
Employment_basis
Location_of_work
Qualification_required_for_the_job

The analytical results highlight that the most important variable impacting the graduate salary is the provider, which accounts for 14.50% of the total impact. This suggests that the reputation and quality of the institution from which a graduate earns their degree can have a significant impact on their earning potential. This is supported by previous research, which has found that graduates from prestigious universities tend to earn higher salaries [62]. The age of the qualifier is the second most important variable as identified in this research, accounting for 9.41% of the total impact. This suggests that the age at which a graduate enters the workforce can have a significant impact on their earning potential. Older graduates tend to earn higher salaries than younger ones. This may be due to a range of factors, including greater work experience and a higher level of seniority in the workplace [63].

The academic year is the third most important variable, accounting for 8.05% of the total impact. Suggesting that the timing of graduation can have an absolute impact on a graduate's earning potential. This may be due to factors such as changes in the labour market, shifts in demand for certain skills, and changes in government policy. The most

important activity is the fourth most important variable, with a feature importance score of 7.65%. This refers to the main activity that a graduate engages in after completing their degree, such as starting a job, further study, or unemployment. This suggests that the transition from education to work is a critical determinant of graduate salaries. Socioeconomic classification is among one of the most important factors impacting graduate salaries in the UK, accounting for 6.34%. The analysis indicates that graduates who come from higher socioeconomic classes, such as higher managerial and professional occupations, tend to have higher salaries than those from lower socioeconomic classes, such as routine and semi-routine occupations. This is aligned with the findings that the socioeconomic status is a significant predictor of educational attainment and subsequent labour market outcomes, including salaries [64]. Individuals from higher socioeconomic classes have more access to resources such as high-quality schools, networks, and mentors that can facilitate educational and career success. In other words, graduates from higher managerial and professional occupations tend to have higher salaries due to their higher levels of education and their access to high-paying jobs and networks. On the other hand, graduates from lower managerial and professional occupations and routine and semi-routine occupations tend to have lower salaries due to the limited opportunities for career advancement and their lower levels of education. Overall, the impact of socioeconomic classification on the graduate salary in the UK is a complex and multifaceted issue that requires further investigation. However, individuals from higher socioeconomic classes tend to have higher salaries than those from lower socioeconomic classes, highlighting the need for policies and interventions that promote equal access to education and career opportunities.

From the analytical result, location of work and domicile are both important variables impacting the graduate salary, accounting for 4.37% and 5.96% respectively. Regarding the impact of location of work, studies have shown that there are significant regional variations in graduate salaries across the UK, which is aligned with the report by the Sutton Trust [65], which found that graduates who work in London tend to earn significantly more than those who work in other regions of the UK. Additionally, a study by the Institute for Fiscal Studies [66] found that graduates who work in London tend to have higher earning trajectories over time compared to those who work in other regions. In terms of domicile, research has found that there are significant differences in graduate earnings depending on where the graduate is from. It is worth noting that some variables, such as ethnicity and disability, were found to have a relatively low impact on the graduate salary. This does not mean that these factors are not important or relevant to the earning potential of graduates, but rather that they may be less significant than other factors in the model.

In inference, the SHAP feature importance model provides valuable insights into the factors that impact graduate salaries in the UK. The results suggest that the reputation and quality of the institution, age of the qualifier, academic year, most important activity, interim study, employment basis, socioeconomic classification, domicile, location of work, publication main activity, and parental education are significant determinants of graduate starting salaries. These findings are supported by previous research and can help inform policy decisions aimed at improving graduate outcomes in the UK.

5.2. Challenges and Limitations of This Study

While machine learning models provide a powerful framework for analysing graduate salary determinants, they have inherent limitations, particularly regarding overfitting and generalisability. Decision trees, despite their interpretability, are prone to memorising training data, making them less reliable when applied to new datasets. Additionally, the resampling techniques used to address class imbalance, such as undersampling and oversampling, may remove valuable data or introduce synthetic noise, affecting model

predictions. Future research should explore more advanced resampling techniques like SMOTE or generative adversarial networks (GANs) to enhance model robustness. Future research should evaluate whether decision tree-based predictions remain consistent across various socioeconomic and geographic contexts.

Beyond model-specific limitations, this study does not account for emerging labour markets trends, such as the rise of online education and AI-driven hiring. Online learning platforms are becoming a viable pathway to career advancement, particularly for working professionals, yet their impact on graduate salaries remains underexplored. Furthermore, with AI-driven automation reshaping job qualifications, many industries are shifting toward skill-based hiring, reducing reliance on traditional degree credentials. Future research should investigate how AI will influence salary determinants and whether degree-based wage advantages will persist in an evolving job market.

5.3. Statistical Significant Interactions Between Contributors and Graduate Salaries

In this study, the variable importance identified by the SHAP model consisting of the top 12 variables that impact the graduate salary in the UK was further investigated. These variables include the provider, academic year, age of qualifier, socioeconomic classification, most important activity, location of work, employment basis, publication main activity, domicile, interim study, qualification required for the job, and parental education. The SHAP model highlighted the percentages by which these variables account for their independent impact on the graduate salary. However, it is not sufficient to stop there. The need to ascertain and investigate if these variables were statistically significant using a well-defined statistical technique was imperative.

Employing the analysis of variance (ANOVA), the variables presented in Table 7 were statistically significant ($p < 0.05$). This result suggests that all the top 12 variables bear significant impacts on the graduate salary. Considering they were all significant, it is evident that both SHAP and ANOVA have individually demonstrated a significant impact on the graduate salary. However, the multi-comparison analysis has revealed statistically significant interactions among the top twelve variables. For instance, it is noteworthy that the provider has the highest percentage impact on the graduate salary; however, it is insufficient to merely attend a prestigious institution without considering the location of work, which has a positive interaction impact with a provider on the salary.

Furthermore, the findings from ANOVA (Table 7) confirm the statistical significance of multiple socioeconomic and demographic factors in determining graduate salaries. However, these results also reveal larger structural patterns in the labour market. The dominance of the institution attended (Provider) as the most significant predictor suggests that graduates from elite universities enjoy a considerable salary advantage, reinforcing existing concerns about higher education elitism and accessibility. While this study identifies a strong link between work location and the salary, particularly favouring London graduates, it does not account for regional living costs, which could diminish real income advantages. Additionally, while the socioeconomic background significantly influences initial salaries, this study does not track whether these disparities persist, decrease, or widen over time. Future research should examine longitudinal earnings data to determine the long-term effects of these factors on career progression.

5.4. Validation of Results by Industry Experts

The validation of results by industry experts is a critical step in research that aims to provide actionable insights and inform evidence-based decision-making [67]. Engaging industry professionals ensures that study findings are reliable, unbiased, and applicable to

real-world scenarios. Additionally, expert validation helps identify variances in findings, contributing to a more comprehensive understanding of the results.

This study administered a structured questionnaire using Google form to industry experts, including seven senior executives from the higher education sector, four top management professionals, six HR representatives, and five other domain-specific specialists. The questionnaire was based on a five-point Likert scale ranging from "strongly disagree" to "strongly agree". This approach enabled a detailed analysis of expert perspectives. The analysis of expert responses, presented in Figure 7, showed that most variables examined in this study were perceived as having a significant impact on graduate salaries, though to varying degrees in the increasing order of agreeing on the factors from top to bottom. Some of the highlights of experts' opinions are presented in Figure 8a,b; the variable "Qualification required for the job" received the highest level of agreement, with approximately 95.24% of experts indicating either "agree" or "strongly agree". The variable "Provider" followed closely, with 90.47% agreement. A pie chart is included to represent these findings visually. The explainability of the machine learning model independently identified "Provider" as the variable with the greatest influence on graduate salaries, aligning closely with expert opinions. Experts also strongly agreed on the importance of "Most important activity" and "Level of study", further supporting the model's findings. This alignment between expert validation and machine learning results reinforces the reliability of this study.

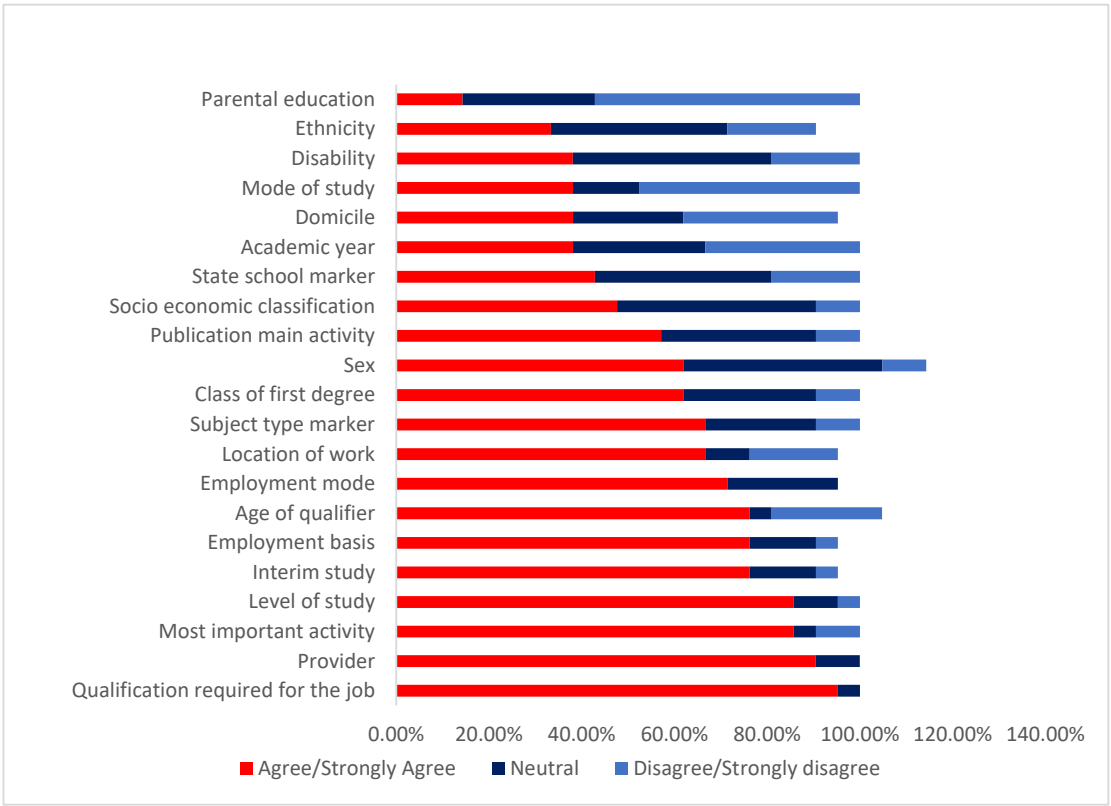

Figure 7. Experts' validation results.

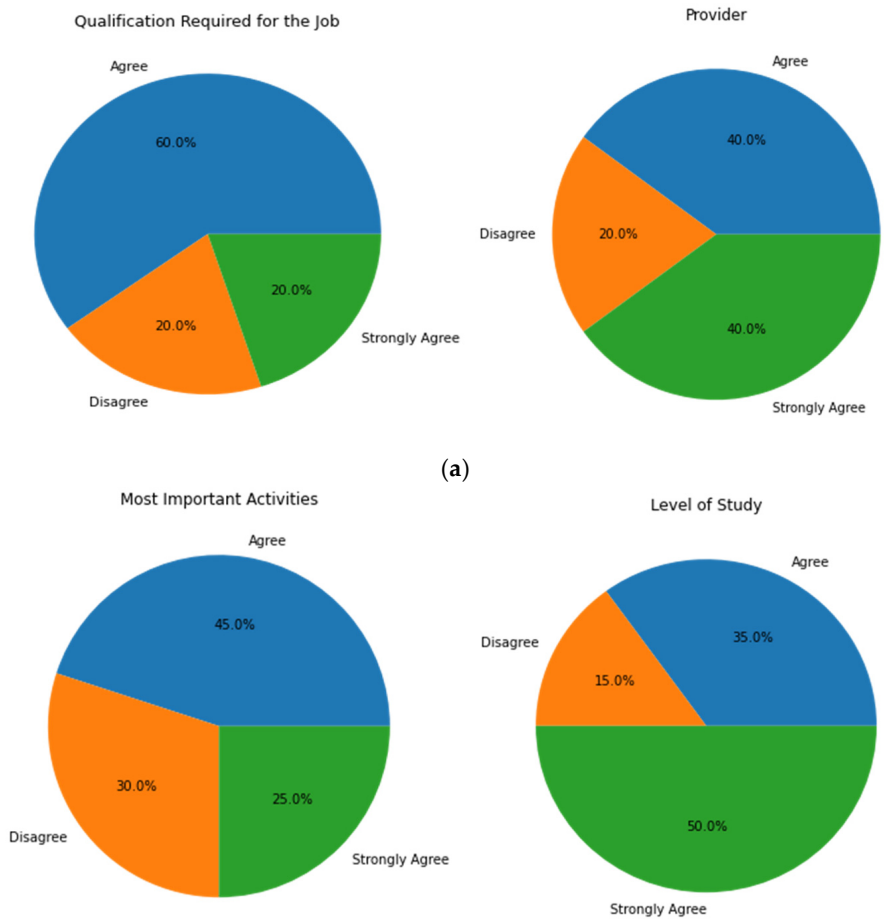

Figure 8. (**a**) Experts' opinion analysis on qualification requirement for the job. (**b**) Experts' opinion analysis on the most important activities.

6. Conclusions and Recommendations

This study bridges the gap between theoretical and practical applications by combining them. This study employed a multi-stage analytical process by integrating ML explanatory and statistical analytical methods with expert responses. The involvement of industry experts, alongside advanced analytical techniques, has produced findings that are credible and applicable to the field of education. This comprehensive validation process enhances this study's relevance and reliability, providing valuable insights for both academic and professional audiences. In conclusion, the analysis of factors influencing graduate salaries in the UK underscores the complex and multifaceted nature of this issue. This study highlights that undersampling is the preferred sampling technique, while decision trees prove to be the most suitable machine learning method for modelling graduate salaries. Additionally, a plethora of factors including the reputation of the institution, age of qualifier, academic year, most important activity, socioeconomic classification, domicile, location of work, parental education, interim study, qualification required for the job, and parental education significantly impact the starting salaries of graduates.

It is noteworthy that this research also reveals binary interaction effects among these variables, further complicating the understanding of graduate salary determinants. These findings are corroborated by unbiased validations from sector experts and previous research. However, it is essential to reflect on the limitations of existing studies, which may have been addressed by this work. Previous research may have lacked comprehensive exploration of

certain factors or failed to consider interaction effects among variables. By addressing these limitations, this study contributes to a more nuanced understanding of graduate salary dynamics in the UK. These insights hold significant implications for policy decisions aimed at enhancing graduate outcomes in the UK. By recognising the intricate interplay of various factors affecting graduate salaries, policymakers can tailor interventions more effectively to support graduates in securing better employment opportunities and higher salaries. Ultimately, this research contributes to the ongoing discourse on improving graduate outcomes and informs evidence-based policy formulation in the UK context.

Based on the analysis findings, it is evident that graduates from higher socioeconomic classes tend to command higher salaries compared to their counterparts from lower socioeconomic backgrounds. For instance, this study revealed that graduates from higher socioeconomic classes had on average, salaries approximately 25.03% higher than those from lower socioeconomic classes. This disparity underscores the need for policies and interventions aimed at promoting equal access to education and career opportunities, particularly for individuals from disadvantaged socioeconomic backgrounds. Furthermore, the analysis highlighted the significant impact of the location of work on graduate salaries with England at 52.64%, Northern Ireland at 14.23%, Scotland at 8.90%, and Wales at 9.45%. Graduates employed in more prosperous areas tend to earn higher salaries compared to those in less prosperous regions. This emphasises the importance of policies promoting regional development and job opportunities in economically disadvantaged areas. By addressing disparities in job availability across regions, such policies could help mitigate the impact of location on graduate salaries, thereby fostering more equitable outcomes for all graduates.

This study's findings have several practical implications for various stakeholders. Higher education institutions should focus on bridging socioeconomic gaps by expanding mentorship programmes, financial aid, and career support services for students from disadvantaged backgrounds. Policymakers must recognise that regional salary disparities and elitism in higher education continue to shape graduate earnings, necessitating targeted economic and educational reforms. Employers, meanwhile, should reassess recruitment strategies to ensure hiring processes emphasise competencies over institutional prestige. Furthermore, the rise of AI and automation could reshape the job market need, meaning continuous evaluation of education strategies covering the need for learning, digital skills, and professional certifications in the near future.

Moreover, while this study identified factors such as parental education as influential determinants of graduate salaries, further research is warranted to explore these factors in greater depth. Despite their importance, parental education was found to have a relatively low impact on graduate salaries in the study model. Therefore, additional research could provide deeper insights into the mechanisms through which parental education influences graduate earnings, informing the design of targeted interventions to support graduates from diverse educational backgrounds. Additionally, there is a need for more comprehensive research into the impact of factors such as ethnicity and disability on graduate salaries. These factors are likely to play significant roles in determining earning potential, yet their influence was not fully explored in the current analysis. By conducting further research in these areas, policymakers can gain a better understanding of the challenges faced by graduates from an ethnic minority or disabled backgrounds and develop tailored strategies to address disparities in salary outcomes.

Overall, the analysis of the factors influencing graduate salaries in the UK provides valuable insights into the complex and multifaceted nature of the issue. The study findings suggest that policies and interventions aimed at promoting equal access to education and career opportunities, as well as regional development, could help to improve graduate

outcomes in the UK. Additionally, further research is needed to investigate the impact of other factors such as parental impacts, location of work, ethnicity and disability on graduate salaries.

Author Contributions: Conceptualisation, Methodology, and Writing—Review and Editing: B.H.; Investigation, Software, and Writing Original Draft: B.H.; Validation: B.H., B.K.M., W.S. and Z.P.; Supervision: B.K.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chankseliani, M. International Development Higher Education: Looking from the Past, Looking to the Future. *Oxf. Rev. Educ.* **2022**, *48*, 457–473. [CrossRef]
2. Green, F.; Henseke, G. Europe's Evolving Graduate Labour Markets: Supply, Demand, Underemployment and Pay. *J. Labour Mark. Res.* **2021**, *55*, 1–13. [CrossRef]
3. Sabiote, C. Use of Binomial Logistic Regression Models for the Study of Determining Variables in the Labour Insertion of University Graduates. *Res. Postgrad.* **2022**, *22*, 109–144.
4. Maquera-Luque, P.J.; Morales-Rocha, J.L.; Apaza-Panca, C.M. Socio-Economic and Cultural Factors That Influence the Labour Insertion of University Graduates, Peru. *Heliyon* **2021**, *7*, e07420. [CrossRef] [PubMed]
5. Walker, I.; Zhu, Y. Impact of University Degrees on the Lifecycle of Earnings: Some Further Analysis. 2013. Available online: https://assets.publishing.service.gov.uk/media/5a7b8cc5e5274a7202e17e36/bis-13-899-the-impact-of-university-degrees-on-the-lifecycle-of-earnings-further-analysis.pdf (accessed on 22 November 2024).
6. Holmes, C.; Mayhew, K. The economics of higher education. *Oxf. Rev. Econ. Policy* **2016**, *32*, 475–496. [CrossRef]
7. Duta, A.; Wielgoszewska, B.; Iannelli, C. Different Degrees of Career Success: Social Origin and Graduates' Education and Labour Market Trajectories. *Adv. Life Course Res.* **2021**, *47*, 100376. [CrossRef]
8. Anders, J. *Does Socioeconomic Background Affect Pay Growth among Early Entrants to High-Status Jobs?* National Institute of Economic and Social Research (NIESR) Discussion Papers; National Institute of Economic and Social Research: London, UK, 2015.
9. Guri-Rosenblit, S.; Šebková, H.; Teichler, U. Massification and Diversity of Higher Education Systems: Interplay of Complex Dimensions. *High. Educ. Policy* **2007**, *20*, 373–389. [CrossRef]
10. Webb, S.; Bathmaker, A.-M.; Gale, T.; Hodge, S.; Parker, S.; Rawolle, S. Higher Vocational Education and Social Mobility: Educational Participation in Australia and England. *J. Vocat. Educ. Train.* **2017**, *69*, 147–167. [CrossRef]
11. Đonlagić, S.; Kurtić, A. The Role of Higher Education in a Knowledge Economy. In *Economic Development and Entrepreneurship in Transition Economies*; Ateljević, J., Trivić, J., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 91–106. ISBN 978-3-319-28855-0.
12. Anderson, R. University Fees in Historical Perspective. History & Policy. 2016. Available online: https://historyandpolicy.org/policy-papers/papers/university-fees-in-historical-perspective (accessed on 22 November 2024).
13. Daly, A.; Lewis, P.; Corliss, M.; Heaslip, T. The Private Rate of Return to a University Degree in Australia. *Aust. J. Educ.* **2015**, *59*, 97–112. [CrossRef]
14. Dietrichson, J.; Bøg, M.; Filges, T.; Klint Jørgensen, A.-M. Academic Interventions for Elementary and Middle School Students with Low Socioeconomic Status: A Systematic Review and Meta-Analysis. *Rev. Educ. Res.* **2017**, *87*, 243–282. [CrossRef]
15. Cooper, D.; Mokhiber, Z.; Zipperer, B. *Raising the Federal Minimum Wage to $15 by 2025 Would Lift the Pay of 32 Million Workers*; Economic Policy Institute (EPI): Washington, DC, USA, 2021; Available online: https://www.epi.org/publication/raising-the-federal-minimum-wage-to-15-by-2025-would-lift-the-pay-of-32-million-workers/ (accessed on 25 November 2024).
16. Bertrand, M.; Duflo, E. Field Experiments on Discriminationa. In *Handbook of Economic Field Experiments*; Elsevier: Amsterdam, The Netherlands, 2017; Volume 1, pp. 309–393. Available online: https://www.sciencedirect.com/science/article/pii/S2214658X1630006X (accessed on 22 November 2024).
17. Blau, F.D.; Kahn, L.M. The Gender Wage Gap: Extent, Trends, and Explanations. *J. Econ. Lit.* **2017**, *55*, 789–865. [CrossRef]
18. Smith, A.E.; Hassan, S.; Hatmaker, D.M.; DeHart-Davis, L.; Humphrey, N. Gender, Race, and Experiences of Workplace Incivility in Public Organizations. *Rev. Public. Pers. Adm.* **2021**, *41*, 674–699. [CrossRef]
19. Chetty, R.; Friedman, J.N.; Saez, E.; Turner, N.; Yagan, D. *Mobility Report Cards: The Role of Colleges in Intergenerational Mobility*; National Bureau of Economic Research: Cambridge, MA, USA, 2017. [CrossRef]
20. Meara, K.; Pastore, F.; Webster, A. The Gender Pay Gap in the USA: A Matching Study. *J. Popul. Econ.* **2020**, *33*, 271–305. [CrossRef]

21. Hogan, R.; Chamorro-Premuzic, T.; Kaiser, R.B. Employability and Career Success: Bridging the Gap Between Theory and Reality. *Ind. Organ. Psychol.* **2013**, *6*, 3–16. [CrossRef]
22. Fugate, M.; Kinicki, A.J. A Dispositional Approach to Employability: Development of a Measure and Test of Implications for Employee Reactions to Organizational Change. *J. Occup. Organ. Psychol.* **2024**, *81*, 503–527. [CrossRef]
23. Rosenberg, S.; Heimler, R.; Morote, E. Basic Employability Skills: A Triangular Design Approach. *Educ. + Train.* **2012**, *54*, 7–20. [CrossRef]
24. Jackson, D.; Sibson, R.; Riebe, L. Delivering Work-Ready Business Graduates—Keeping Our Promises and Evaluating Our Performance. *J. Teach. Learn. Grad. Employab.* **2013**, *4*, 2–22. [CrossRef]
25. Cai, Y. Graduate Employability: A Conceptual Framework for Understanding Employers' Perceptions. *High. Educ.* **2013**, *65*, 457–469. [CrossRef]
26. Light, A.; Strayer, W. Who receives the college wage premium?: Assessing the labor market returns to degrees and college transfer patterns. *J. Hum. Resour.* **2004**, *39*, 746–773. [CrossRef]
27. Zhang, L. Gender and racial gaps in earnings among recent college graduates. *Rev. High. Educ.* **2008**, *32*, 51–72. [CrossRef]
28. Taniguchi, H. The influence of age at degree completion on college wage premiums. *Res. High. Educ.* **2005**, *46*, 861–881. [CrossRef]
29. Quadlin, N.; VanHeuvelen, T.; Ahearn, C.E. Higher Education and High-Wage Gender Inequality. *Soc. Sci. Res.* **2023**, *112*, 102873. [CrossRef] [PubMed]
30. Davidovitch, N. Discipline and Gender: Factors Affecting Graduates' Salaries. *US-China Foreign Lang.* **2013**, *11*, 770–778. [CrossRef]
31. Macmillan, L.; Tyler, C.; Vignoles, A. Who Gets the Top Jobs? The Role of Family Background and Networks in Recent Graduates' Access to High-Status Professions. *J. Soc. Pol.* **2015**, *44*, 487–515. [CrossRef]
32. Charizanos, G.; Demirhan, H.; İçen, D. Binary Classification with Fuzzy Logistic Regression under Class Imbalance and Complete Separation in Clinical Studies. *BMC Med. Res. Methodol.* **2024**, *24*, 145. [CrossRef]
33. Cunningham, P.; Delany, S.J. K-Nearest Neighbour Classifiers—A Tutorial. *ACM Comput. Surv.* **2022**, *54*, 1–25. [CrossRef]
34. Alizadeh, E.; Maleki, A. A Novel K-NN Algorithm for Imbalanced Datasets Using Euclidean-Smote. *Eng. Appl. Artif. Intell.* **2021**, *100*, 104163. [CrossRef]
35. López-Meneses, E.; López-Catalán, L.; Pelícano-Piris, N.; Mellado-Moreno, P.C. Artificial Intelligence in Educational Data Mining and Human-in-the-Loop Machine Learning and Machine Teaching: Analysis of Scientific Knowledge. *Appl. Sci.* **2025**, *15*, 772. [CrossRef]
36. Wang, H.; Zhang, Y.; Zou, B.; Lui, H.; Wei, X. An Improved Linear Discriminant Analysis Based on Pseudo Inverse for Small Sample Size Problem. *Neural Comput. Appl.* **2021**, *33*, 3931–3939.
37. Ma, S.; Huang, H.; Luo, Y.; Gao, Y. Feature Selection Based on Improved Fisher's Linear Discriminant Analysis for Fault Diagnosis of Rotating Machinery. *Mech. Syst. Signal Process* **2021**, *157*, 107752.
38. Safi, S.K.; Gul, S. An Enhanced Tree Ensemble for Classification in the Presence of Extreme Class Imbalance. *Mathematics* **2024**, *12*, 3243. [CrossRef]
39. Lu, J.; Wu, H.; Wu, Y.; Wang, J.; Zhou, Y. An Improved Decision Tree Algorithm Based on Information Gain Ratio. *J. Intell. Fuzzy Syst.* **2021**, *40*, 41–50.
40. Atzmueller, M.; Fürnkranz, J.; Kliegr, T.; Schmid, U. Explainable and Interpretable Machine Learning and Data Mining. *Data Min. Knowl. Disc* **2024**, *38*, 2571–2595. [CrossRef]
41. Mohri, M.; Rostamizadeh, A.; Talwalkar, A. *Foundations of Machine Learning*, 2nd ed.; Adaptive Computation and Machine Learning; MIT Press: Cambridge, MA, USA, 2018; ISBN 978-0-262-03940-6.
42. Lee, H.; Jang, Y.; Lee, J.; Kim, J. Robust Decision Tree Pruning via Rule-Based Regularization. *Neural Netw.* **2021**, *144*, 12–23. [CrossRef]
43. Fuseini, I.; Missah, Y.M. A Critical Review of Data Mining in Education on the Levels and Aspects of Education. *Qual. Educ. All* **2024**, *1*, 41–59. [CrossRef]
44. Han, S.; Kim, H.; Lee, Y.-S. Double Random Forest. *Mach. Learn.* **2020**, *109*, 1569–1586. [CrossRef]
45. Shashaani, S.; Sürer, Ö.; Plumlee, M.; Guikema, S. Building Trees for Probabilistic Prediction via Scoring Rules. *Technometrics* **2024**, *66*, 625–637. [CrossRef]
46. Hue, C.; Boullé, M.; Lemaire, V. Online Learning of a Weighted Selective Naive Bayes Classifier with Non-Convex Optimization. In *Advances in Knowledge Discovery and Management*; Guillet, F., Pinaud, B., Venturini, G., Eds.; Studies in Computational Intelligence; Springer International Publishing: Cham, Switzerland, 2017; Volume 665, pp. 3–17; ISBN 978-3-319-45762-8.
47. Khan, A.A.; Chaudhari, O.; Chandra, R. A Review of Ensemble Learning and Data Augmentation Models for Class Imbalanced Problems: Combination, Implementation and Evaluation. *Expert. Syst. Appl.* **2024**, *244*, 122778. [CrossRef]
48. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *arXiv* **2016**, arXiv:1603.02754. [CrossRef]
49. Sohil, F.; Sohali, M.U.; Shabbir, J. An Introduction to Statistical Learning with Applications in R. *Stat. Theory Relat. Fields* **2022**, *6*, 87. [CrossRef]

50. Cervantes, J.; Garcia-Lamont, F.; Rodríguez-Mazahua, L.; Lopez, A. A Comprehensive Survey on Support Vector Machine Classification: Applications, Challenges and Trends. *Neurocomputing* **2020**, *408*, 189–215. [CrossRef]
51. Seo, Y.; Shin, K. Hierarchical Convolutional Neural Networks for Fashion Image Classification. *Expert. Syst. Appl.* **2019**, *116*, 328–339. [CrossRef]
52. García-Blanco, M.; Cárdenas Sempértegui, E.B. Labour Insertion in Higher Education: The Latin American Perspective. *Educ. XX1* **2018**, *21*, 323–347. [CrossRef]
53. Delaney, B.; Tansey, K.; Whelan, M. Satellite Remote Sensing Techniques and Limitations for Identifying Bare Soil. *Remote Sens.* **2025**, *17*, 630. [CrossRef]
54. Ren, Y.; Xie, Z.; Zhai, S. Urban Land Use Classification Model Fusing Multimodal Deep Features. *ISPRS Int. J. Geo-Inf.* **2024**, *13*, 378. [CrossRef]
55. Bengio, S.; Deng, L.; Larochelle, H.; Lee, H.; Salakhutdinov, R. Guest Editors' Introduction: Special Section on Learning Deep Architectures. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1795–1797. [CrossRef]
56. Kiran, R.J.; Sanil, J.; Asharaf, S. A Novel Approach for Model Interpretability and Domain Aware Fine-Tuning in AdaBoost. *Hum-Cent. Intell. Syst.* **2024**, *4*, 610–632. [CrossRef]
57. James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning: With Applications in R*, 2nd ed.; Springer Texts in Statistics; Springer: New York, NY, USA, 2021; ISBN 978-1-07-161418-1.
58. Curley, C.; Krause, R.M.; Feiock, R.; Hawkins, C.V. Dealing with Missing Data: A Comparative Exploration of Approaches Using the Integrated City Sustainability Database. *Urban. Aff. Rev.* **2019**, *55*, 591–615. [CrossRef]
59. Chen, W.; Yang, K.; Yu, Z.; Shi, Y.; Chen, C.L.P. A Survey on Imbalanced Learning: Latest Research, Applications and Future Directions. *Artif. Intell. Rev.* **2024**, *57*, 137. [CrossRef]
60. Rudin, C.; Chen, C.; Chen, Z.; Huang, H.; Semenova, L.; Zhong, C. Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges. *Statist. Surv.* **2022**, *16*, 1–85. [CrossRef]
61. Salih, A.M.; Raisi-Estabragh, Z.; Galazzo, I.B.; Radeva, P.; Petersen, S.E.; Lekadir, K.; Menegaz, G. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Adv. Intell. Syst.* **2025**, *7*, 2400304. [CrossRef]
62. Zhang, B.; Zhang, Q.; Yao, C.; Liu, Z. The Signaling Paradox: Revisiting the Impacts of Overeducation in the Chinese Labour Market. *Educ. Sci.* **2024**, *14*, 900. [CrossRef]
63. Asfaw, A. Racial and Ethnic Disparities in Teleworking Due to the COVID-19 Pandemic in the United States: A Mediation Analysis. *Int. J. Environ. Res. Public Health* **2022**, *19*, 4680. [CrossRef] [PubMed]
64. McKnight, A.; Naylor, R. Going to University: The Influence of Schools, Information and Parental Expectations. *Oxf. Rev. Educ.* **2015**, *41*, 231–248.
65. Sutton, T. Mobility Manifesto: A Lifetime of Opportunities. *Mobil. Manif.* **2020**. Available online: https://www.suttontrust.com/wp-content/uploads/2019/12/Mobility-Manifesto-2019-1.pdf (accessed on 24 November 2024).
66. Institute for Fiscal Studies. The Impact of Higher Education on Regional Economies and Lifetime Earnings. IFS. 2020. Available online: https://ifs.org.uk/publications/impact-undergraduate-degrees-lifetime-earnings (accessed on 25 November 2024).
67. Pérez-Rivas, F.J.; Jiménez-González, J.; Bayón Cabeza, M.; Belmonte Cortés, S.; De Diego Díaz-Plaza, M.; Domínguez-Bidagor, J.; García-García, D.; Gómez Puente, J.; Gómez-Gascón, T. Design and Content Validation Using Expert Opinions of an Instrument Assessing the Lifestyle of Adults: The 'PONTE A 100' Questionnaire. *Healthcare* **2023**, *11*, 2038. [CrossRef]