International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

Check for updates

From academia to industry: A framework to securely implement big data and AI to predict college graduates' employment trajectories

Muhammad Faizan [1,*], Qiming Huang [1], Nayab Riaz [2] and Usman Saif [1]

[1] *School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing, P.R. China.*
[2] *School of Management Science and Engineering, University of Science and Technology Beijing, Beijing, P.R. China.*

International Journal of Science and Research Archive, 2024, 11(02), 708–723

Publication history: Received on 13 February 2024; revised on 22 March 2024; accepted on 25 March 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.11.2.0497

Abstract

The transition from academia to industry can be unpredictable, but what if we could forecast college graduate employment outcomes with both accuracy and robust security? This study introduces an innovative framework that leverages secure data analysis and machine learning to predict the employment trajectories of college graduates. By integrating homomorphic encryption, we safeguard the privacy of sensitive personal and academic data while enabling complex machine learning operations. Our approach involves meticulous data collection, feature engineering, encryption, and model development, resulting in a robust model that addresses privacy concerns without sacrificing prediction accuracy. We demonstrate our model's superiority over traditional approaches, achieving a notable increase in both security and stability. This research illuminates the potential of encrypted data analysis in reshaping predictive modeling methods, offering insights for educational institutions, policymakers, and students. Our findings not only address a pressing issue in employment forecasting but also lay the groundwork for secure and ethical big data applications across various domains.

Keywords: Secure Data Analytics; Privacy-Preserving AI; Homomorphic Encryption; Employment Forecasting; Educational Analytics

1. Introduction

The transition from college to the workforce is fraught with uncertainty. Imagine a recent graduate, armed with a degree and aspirations, facing a job market in constant flux. Traditional employment prediction tools, often relying on limited historical data and basic statistical models, often provide inadequate guidance (1). This leaves students and policymakers feeling ill-equipped to navigate the complexities of the modern job market. The rise of big data and machine learning offers a glimmer of hope, but how can we harness these technologies while upholding the fundamental right to privacy? This study proposes an innovative solution – a predictive model that merges the power of big data with the security of homomorphic encryption to accurately forecast college graduate employment outcomes while safeguarding individual privacy. In the evolving landscape of employment, accurately predicting the career trajectories of college graduates has emerged as a critical challenge. While foundational, traditional methodologies often fall short in addressing the complexities of a job market marked by rapid technological advancements and fluctuating economic conditions (2). These conventional approaches struggle to capture the multifaceted nature of employment trends and the interplay between individual skills and labor market demands. As a result, there's a pressing need for innovative solutions that not only enhance the precision of predictions but also navigate the delicate balance of data privacy. The advent of big data technology and machine learning heralds a new era in employment prediction. These technologies offer the potential to process and analyze vast volumes of diverse data, from academic records and personal achievements to market trends and industry demands (3). By leveraging such comprehensive datasets, researchers and policymakers can uncover intricate patterns and correlations that were previously obscured. However, this approach

* Corresponding author: Muhammad Faizan

introduces a new set of challenges, chief among them the concern for the privacy and security of individuals' information. The concept of homomorphic encryption emerges as a beacon of hope in this context. It promises a way to utilize sensitive data for predictive analytics without compromising privacy, thus aligning with the growing emphasis on ethical data usage (4). This study aims to explore the intersection of these advanced technologies—how they can be harnessed to not only predict employment outcomes with accuracy but do so in a manner that safeguards the personal information of graduates.

This sets the stage for a comprehensive exploration into developing a predictive model that embodies the strengths of big data analytics and machine learning while adhering to the principles of data privacy through homomorphic encryption. The significance of this research extends beyond academic interest; it has the potential to influence educational policy, curriculum development, and individual career planning, ultimately contributing to a more informed and responsive education system. In light of the evolving job market and the pressing need for enhanced predictive accuracy and data privacy, this study seeks to address the gaps left by traditional graduate employment prediction methods. By embracing the capabilities of big data, machine learning, and encryption technologies, this research aspires to offer novel insights and tools that can adapt to the dynamic nature of the job market while prioritizing the privacy of individuals' data. This endeavor not only addresses an immediate analytical challenge but also reflects a broader commitment to ethical and responsible data use in the digital age. Therefore, this multifaceted approach necessitates a clear set of objectives to guide the research process: (a) To design and validate a predictive model enhanced by big data analytics for accurately forecasting employment rates among college graduates. (b) To integrate homomorphic encryption into the predictive model, ensuring the privacy and security of sensitive data throughout the analysis process. (c) To assess the effectiveness of the proposed model against traditional employment prediction methodologies, highlighting improvements in performance and data security. (d) To explore the implications of the findings for policy-making and educational strategy development, particularly in terms of curriculum adjustments to better prepare students for the job market. (e) To contribute to the broader discourse on ethical data usage in predictive analytics, setting a precedent for future research in secure data analysis.

The significance of this study lies in its innovative approach to maintaining the predictive accuracy of college graduates' employment outcomes while ensuring data privacy through homomorphic encryption. It represents a meaningful advancement in the application of big data and machine learning in the educational sector, with potential implications for policy-making and curriculum development. By offering insights into secure, privacy-preserving data analysis, this research could set new standards for handling sensitive information in predictive modeling, thereby influencing future methodologies in various fields beyond education. The expected results and findings promise to contribute valuable knowledge to the academic community and offer practical guidance for improving graduate employment rates.

2. Literature Review

As data-driven decision-making permeates various sectors, data privacy and security concerns escalate, including in human resources and employment prediction. Predictive models using sensitive data like academic performance, work experience, or personality assessments offer immense potential for talent identification but raise critical privacy questions. Regulations like the GDPR (5) and CCPA (6) mandate secure data handling practices and grant individuals control over their personal information. Traditional analytics methods, requiring cleartext data access, pose risks of unauthorized access or breaches. Recent university data breach exposed student records, including GPA and test scores (7), emphasizing need for secure analysis techniques to protect individual privacy while enabling valuable insights.

A vast body of research highlights the link between academic performance and employment outcomes, consistently demonstrating a positive correlation between higher GPAs and securing desirable jobs (8). (9), for example, found that graduates with higher GPAs enjoyed more job offers and higher salaries. This likely reflects how academic success signals valuable skills like strong work ethic, self-discipline, and problem-solving. However, recent studies emphasize the limitations of relying solely on academic achievement. (10) stress the rising demand for soft skills like communication, teamwork, and adaptability. Industry-specific knowledge and relevant work experience are also crucial. These findings point to the need for predictive models that broaden their scope beyond traditional academic metrics, integrating transferable skills, internships, and extracurricular activities for enhanced accuracy. Technological advancements are rapidly transforming the job market. Burning Glass Technologies (11) reveals a surge in demand for data analysis, machine learning, and programming skills across industries. This shift highlights a skills gap between graduates and employer needs. The World Economic Forum (12) predicts the emergence of up to 97 million new jobs by 2030, emphasizing the importance of artificial intelligence and big data skills. To thrive in this landscape, educational institutions must adapt their curricula to equip graduates with the necessary skillset. Research underscores the correlation between employability and in-demand skills. (13) demonstrate that individuals with strong analytical and data-driven skills are more likely to secure employment and promotions. Similarly, (14) highlight the importance of

continuous learning throughout one's career path. Equipping students with these skills requires integrating data analysis and programming courses into curricula alongside fostering work-integrated learning opportunities. Alumni networks are valuable career development resources. Studies demonstrate their significant contributions to job placement and guidance (15). These networks connect alumni with industry professionals, offering mentorship, job postings, and insights. Network effectiveness can vary based on industry, university prestige, and structure, with specialized fields often seeing greater benefits (16). Universities with active alumni outreach and online platforms often achieve higher graduate employment rates. Technology further enhances alumni networks, with online platforms, mobile apps, and AI-powered recommendation systems facilitating connections and personalized career guidance (15).

Homomorphic encryption (HE) emerges as a promising solution for secure predictive modeling. It allows computations directly on encrypted data, ensuring privacy even in untrusted environments (17). Partially homomorphic encryption (PHE) supports specific operations, while fully homomorphic encryption (FHE) enables arbitrary computations. HE empowers organizations to comply with data privacy regulations while fostering collaboration, as sensitive data can be securely analyzed by third-party experts without revealing its underlying content. Applying HE for secure employment prediction is a promising but emerging field. Studies like (18) demonstrate the feasibility of using FHE for secure decision trees in applicant screening, preserving data privacy. A collaboration between Microsoft and Carnegie Mellon University (19) further showcases HE's potential in creating secure machine learning frameworks for analyzing resumes without compromising confidentiality. While HE offers potential for secure and ethical talent evaluation, challenges like computational complexity remain (17). Ongoing research aims to develop more efficient FHE schemes to address this limitation. As secure predictive modeling evolves, addressing ethical concerns like algorithmic bias is crucial to prevent perpetuating discrimination (20). Organizations must prioritize research into fair AI models while staying updated on evolving regulations around data privacy and algorithmic fairness to ensure compliance.

This literature review explored the increasing demand for secure predictive modeling within employment prediction. Homomorphic encryption (HE) offers a promising solution for preserving individual privacy while analyzing sensitive data. Challenges and opportunities associated with HE-based techniques were highlighted, emphasizing the need to improve efficiency and scalability. The importance of ethical considerations and regulatory compliance in developing secure predictive models was stressed. As HE technology matures, alongside other privacy-preserving techniques, it holds the potential to transform data analysis across various fields. However, rigorous research into the accuracy, scalability, and potential biases of HE-based models remains crucial for responsible adoption.

3. Methodology

3.1. Research Design

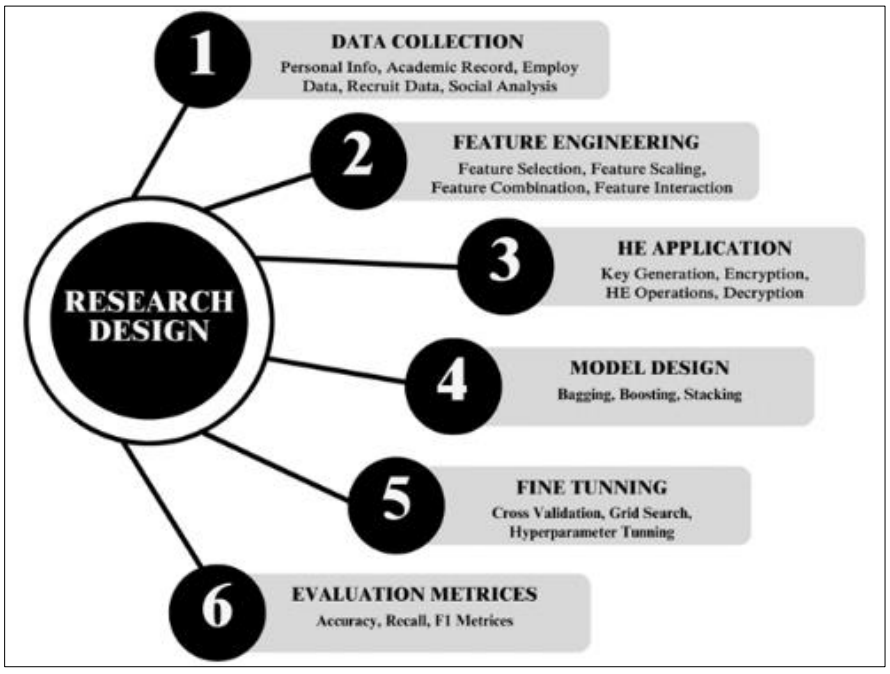

Figure 1 Framework to securely predict graduates' employment using big data and AI

This study adopts a multidisciplinary approach to predict graduate employment outcomes, integrating big data analytics, machine learning, and homomorphic encryption (HE) for enhanced predictive power and data privacy. Our methodology involves comprehensive dataset collection, meticulous feature engineering, and the application of state-of-the-art encrypted machine learning models (Fig 1). This design addresses the dual challenges of boosting prediction accuracy while safeguarding sensitive student information in the digital age. We anticipate this approach will advance employment prediction models and promote ethical data analysis in education. The step-by-step flow of proposed framework is represented in (Fig 2).

3.2. Data Collection and Processing

To build a robust graduate employment prediction model, we collect diverse datasets crucial for understanding factors influencing employment outcomes. This meticulous approach ensures our model is both accurate and reflects the real-world complexities of the job market.

- ***Personal Information Collection:*** From university records, we collect gender, age, ethnicity, and domicile. This data allows us to analyze how diverse backgrounds might influence career paths and ensure the model accommodates these potential variations.
- ***Academic Performance Acquisition:*** Academic records provide grades, credits, and GPA, offering insights into academic aptitude and dedication. This data is essential for correlating academic achievement with employment success.
- ***Employment-Related Data Collection:*** University surveys yield information on internships, social practices, honors, and awards. This sheds light on graduates' practical experience and overall suitability for the workforce – factors that significantly impact employability.
- ***Recruitment Website and Job Information Crawling:*** Using advanced web crawling, we extract job positions, salary ranges, and required skills from recruitment websites. This provides a real-time snapshot of the job market, ensuring our model aligns with current trends and employer expectations.
- ***Social Network Analysis:*** We employ network analysis on social media platforms to uncover relationships and activities among graduates, revealing patterns in career trajectories and potential influential factors. This innovative approach delves into how social dynamics and online interactions shape career development.

Table 1 Data collection overview for graduate employment prediction study

Data Category	Data Types	Sources	Relevance
Personal Information	Gender, Age, Ethnicity, Domicile	University Records	Helps understand the diverse backgrounds of graduates
Academic Performance	Grades, Credits, GPA	Academic Records	Indicates academic abilities and performance
Employment-Related Data	Internship Experience, Social Practice	University Surveys	Reflects practical experience and comprehensive qualities
Recruitment Website Data	Job Positions, Salary Levels	Web Crawling Recruitment Sites	Provides real-time job market information
Social Network Analysis	Relationships, Activities	Social Media Platforms	Offers insights into graduates' communication and decision-making patterns

Following collection, we undertake a rigorous data processing protocol to ensure accuracy and relevance for analysis.

- ***Data Cleaning:*** Data is meticulously cleansed to address inconsistencies, errors, and missing values, enhancing data quality.
- ***Feature Extraction:*** The most predictive features influencing employment outcomes are carefully extracted from the diverse datasets.
- ***Data Integration:*** Disparate datasets are seamlessly integrated, creating a comprehensive dataset primed for developing our predictive model.

This methodical data collection and processing approach strengthens our model's accuracy, reliability, and applicability, setting a high standard for employment prediction analytics.

3.3. Feature Engineering and Selection

In employment prediction, meticulous feature engineering and selection are crucial for building an accurate model that reflects real-world outcomes. By carefully analyzing and selecting the most influential features, we significantly enhance the model's predictive power. This process ensures our predictions offer practical, actionable insights that can guide students towards successful career paths, a key goal of our framework.

3.3.1. Feature Selection

To ensure our graduate employment prediction model is both accurate and reflects real-world factors, meticulous feature selection is critical. We employ a combination of practical considerations, statistical analysis, and feature selection algorithms to identify the variables that have the greatest influence on employment outcomes.

- ***Chi-squared Test:*** This test provides a rigorous statistical method to assess the relationship between categorical variables and the target variable (i.e., employment outcomes). It uses the formula $\chi^2 = \Sigma(fo - fe)^2 / fe$, where '$fo$' is the observed frequency and 'fe' is the expected frequency under the hypothesis of independence. A high Chi-squared value signifies a strong correlation, indicating features worth retaining for further analysis (21).
- ***Mutual Information:*** We calculate the mutual information between each feature and the target variable using: $(I(x;y) = \sum\sum p(x,y) \cdot log(\frac{p(x,y)}{p(x) \cdot p(y)})$. In this context, (x) and (y) represent two random variables, with $(p(x,y))$ denoting the joint probability of (x) and (y) occurring together, $(p(x))$ and $(p(y))$ their individual probabilities. Mutual information quantifies the dependency between variables, with higher values signifying a strong association with employment outcomes (22).
- ***L1 Regularization:*** To avoid overfitting and enhance model interpretability, we employ L1 regularization. This technique penalizes model complexity, effectively pushing less influential features towards zero. The equation is: $(J(\theta) = \text{Loss}(\theta) + \lambda \cdot \sum|\theta|), where(J(\theta))$ represents the regularized loss function, (λ) the initial loss function, $(\text{Loss}(\theta))$ the regularization strength, and $(\text{Loss}(\theta))$ the L1 norm of the parameter vector. By tuning the regularization strength (λ), we control the sparsity of feature weights, ensuring a streamlined model focused on the most impactful predictors (23).

Throughout our feature selection process, we leverage insights from our diverse datasets (e.g., academic and social network data) alongside expert knowledge in career development. This ensures that the final features chosen are statistically sound and align with the practical realities of the job market.

3.3.2. Feature Scaling

To optimize machine learning model performance, we employ feature scaling. This technique addresses discrepancies in feature value ranges, enhancing model stability and convergence speed (24). We either standardize features (mean of 0, variance of 1), or normalize them within a [0,1] range. This preprocessing step is crucial for ensuring all features are treated equitably by the model. The standardization process adjusts each feature value (x) by subtracting the mean (μ) and dividing by the standard deviation $(\sigma), (x' = \frac{x - \mu}{\sigma})$. Meanwhile, normalization adjusts (x) to $(x' = \frac{x - \min(x)}{\max(x) - \min(x)})$, where $(\min(x))$ and $(\max(x))$ are the minimum and maximum values of the feature, respectively. These methods enhance model stability and convergence speed.

3.3.3. Feature Combination and Interaction

Recognizing that graduate employment outcomes are influenced by complex interactions, we carefully engineer new features to capture these nuances. We combine existing features using ratios, differences, and polynomial terms to uncover non-linear relationships. Additionally, feature crossing examines interactions between feature pairs (e.g., multiplying age and GPA), allowing the model to identify complex patterns that individual features might fail to capture. This approach highlights the multi-faceted determinants of employment success, enhancing our predictive capability.

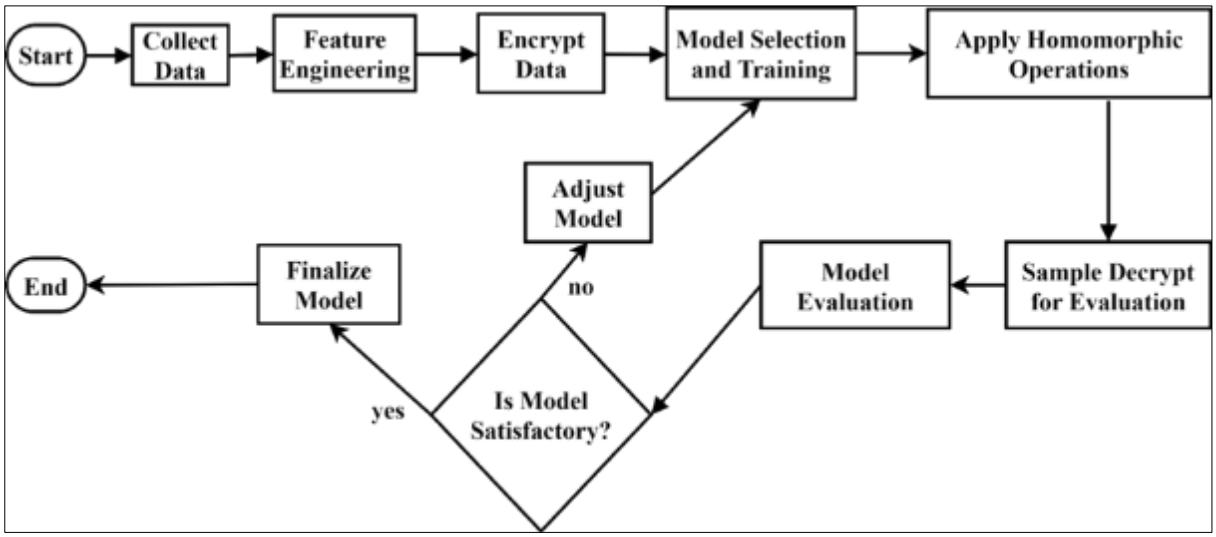

Figure 2 Flow diagram to securely predict graduates' employment using big data and AI

3.4. Homomorphic Encryption Application

To safeguard sensitive student data while enabling robust analysis, we apply homomorphic encryption (HE) after the feature engineering phase. This technique transforms the data, allowing computations to be performed directly on the encrypted values. This approach is essential for upholding strict privacy standards while deriving meaningful insights – a key principle of our framework (25).

3.4.1. Data Encoding and Key Generation

To facilitate secure computations using homomorphic encryption (HE), we begin by carefully transforming employment data into a suitable format. Raw data is encoded numerically for HE operations. Numerical data is converted using binary encoding, while categorical data undergoes one-hot encoding (each category becomes a binary vector with one active element). This encoded data is represented as (x) and mapped to a polynomial $(p(x))$ for further processing.

To ensure secure encryption, we select two large prime numbers (p) and (q), choose two polynomials $(a(x))$ and $(s(x))$ within a specific mathematical structure called a ring. The polynomial $(s(x))$ has small coefficients and $(a(x))$ remains the secret key. This process generates: Public Key: $((p,q,a(x)))$ used for encrypting data. Private Key: $(s(x))$ used for decryption.

3.4.2. Encryption

Following the encoding process, we encrypt the data using the public key. The encryption formula, $(c = p(x)^e \mod n)$, combines the plaintext t polynomial $(p(x))$ with the public key (e) and a small-coefficient noise polynomial (n), yielding the ciphertext (c). This intricate process encapsulated in equation $(C = p(x) + q \cdot a(x) \cdot e(x))$ effectively obfuscates the original data, shielding it from unauthorized access while preserving the ability to perform computations on the encrypted data.

3.4.3. Homomorphic Operations

The true power of homomorphic encryption (HE) lies in its ability to perform calculations directly on encrypted data. This unique property means that when we decrypt the results of these operations, they mirror the outcomes we would get if we performed the same calculations on the original, unencrypted data. Mathematically, if a function (f) is applied to ciphertext (c) resulting in $(c' = f(c))$, upon decryption (c') we will obtain the same result as if we had applied (f) directly to the plaintext (p), hence $(p' = f(p))$.

HE supports various operations. For example, adding two ciphertexts (C_1) and (C_2) is equivalent to adding their underlying plaintext values.

$$[C_{\text{result}} = C_1 + C_2 = (p_1(x) + q \cdot a(x) \cdot e_1(x)) + (p_2(x) + q \cdot a(x) \cdot e_2(x)) = p_1(x) + p_2(x) + q \cdot a(x) \cdot (e_1(x) + e_2(x))]$$

Multiplication on ciphertexts (C_1) and (C_2) is more complex, involving polynomial multiplication within the defined mathematical ring.

This capability unlocks secure analysis on sensitive data. We can perform calculations needed for our employment prediction model without ever exposing the original student information. HE introduces computational overhead. Complexity of operations, especially multiplication, requires considering algorithm efficiency and resource allocation.

3.4.4. Decryption

To evaluate our model's performance on encrypted data, we selectively decrypt a portion of the results using the private key $(s(x))$ undergoing modular reduction by prime (q). The decryption algorithm, $(p(x) = (C \cdot s(x)) \mod q)$, essentially reverses the encryption process, converting the ciphertext (C) back into its original plaintext form (p). This step is crucial for understanding model's predictions and ensuring its results align with real-world employment trends.

3.4.5. Decoding

After decrypting the results of our homomorphic computations, we must decode them to obtain meaningful insights. Decoding reverses any encoding and scaling processes applied to the data before encryption. This step converts the decrypted data $(p(x))$ back into its original numerical or categorical format (e.g., salary ranges, job titles). This transformation is essential for interpreting the model's predictions and ensuring the results are actionable for stakeholders such as educators and policymakers who rely on these insights for decision-making.

3.5. Model Selection and Design

Model selection is especially critical when working with homomorphically encrypted data. The computational overhead of HE mandates careful consideration of model complexity and its potential impact on performance. Linear and logistic regression models are often preferred due to their relative simplicity and lower computational demands (25). For greater accuracy, we may explore ensemble methods like Bagging and Boosting. These techniques combine multiple weaker learners to create a more robust model, potentially improving predictive capabilities while working within the constraints of encrypted data analysis.

- ***Bagging:*** To boost model performance on encrypted data, we consider Bagging (Bootstrap Aggregating). This ensemble technique reduces overfitting and improves accuracy by training multiple models on different subsets of the dataset and then combining their predictions (often through averaging or voting) (26). The aggregation of predictions from (m) models for (n) samples is mathematically represented as $(Y_j = \sum_{i=1}^{m} \frac{Y_i}{m})$, where (Y_i) denotes the prediction of the (i^{th}) model, and (Y_j) signifies the aggregated prediction for the (j^{th}) sample. Bagging is computationally efficient, making it suitable for HE where complexes models are less feasible.
- ***Boosting:*** Boosting is another powerful ensemble method that can improve prediction accuracy, especially with homomorphically encrypted data (27). It tackles misclassified data points iteratively, assigning weights to each model in the ensemble based on its performance. AdaBoost, the predictive power is aggregated as shown: $[Y_j = \sum_{i=1}^{m} (\omega_i \cdot Y_i)]$, where (ω_i) represents the weight of the (i^{th}) model, influencing the overall prediction (Y_j) for each sample. Boosting offers the potential for increased accuracy and adaptability, making it a valuable technique for complex datasets where individual models might struggle.
- ***Stacking:*** To further enhance model performance, we may consider Stacking. This sophisticated ensemble technique combines predictions from multiple diverse models in a hierarchical structure. Outputs of one layer of models become the inputs for the next layer, allowing the meta-learner to identify complex patterns and refine predictions (28). Stacking's ability to blend strengths of different algorithms makes it well-suited for complex tasks like employment prediction, particularly when working with homomorphically encrypted data.

Feature engineering and model design are iterative processes, demanding continuous refinement based on real-world data. To ensure the model's reliability and interpretability, we must prioritize its explanatory power and conduct thorough validation procedures.

3.6. Fine Tunning

Fine-tuning is essential for optimal model performance, especially with complex tasks like those involving homomorphically encrypted data. This process employs techniques like cross-validation, hyperparameter tuning, and grid search to carefully adjust model parameters. By optimizing these settings, we enhance the model's ability to generalize to unseen data while minimizing overfitting.

- ***Cross Validation:*** To rigorously evaluate our model's ability to generalize to unseen data, we employ cross-validation. This technique partitions the dataset into *(k)* subsets (or folds) (29). The model is trained on *k-1* folds and validated on the remaining fold. This process rotates through each fold, yielding *k* performance scores. The cross-validation score, often an average across these scores, calculated as $(CV_k = \frac{1}{k}\sum_{i=1}^{k} ModelScore(M_i, D_i))$, provides a robust estimate of the model's true performance, helping to combat overfitting. (CV_k) is the cross-validation score over *(k)* iterations, (M_i) is the model trained on all data except the (i^{th}) fold, and (D_i) is the (i^{th}) fold of data used for validation.
- ***Hyperparameter Tunning:*** Hyperparameters are settings that control a model's learning process, distinct from the parameters learned during training. Finding the optimal hyperparameter values is crucial for maximizing model performance (30). This often involves a validation process where different combinations are evaluated, aiming to minimize a loss function, $[L(H) = \sum_{i=1}^{n} LossFunction(M(H), D_i)]$ where *(L)* is the loss over the dataset for hyperparameters *(H), (M)*, (D_i) is the model, and is the training data. Effective hyperparameter tuning can lead to a model that converges faster and generalizes better to new data.
- ***Grid Search:*** Grid search provides a systematic way to explore the hyperparameter space. It evaluates the model on all possible combinations of hyperparameters within a predefined grid (31). This exhaustive approach helps identify the combination that yields optimal performance. If your hyperparameter space includes values for k different hyperparameters $[Grid = \{(h_1^1, h_2^1, \ldots, h_m^1), (h_1^2, h_2^2, \ldots, h_m^2), \ldots, (h_1^p, h_2^p, \ldots, h_m^p)\}]$ where (h_j^i) is the (i^{th}) value of the (j^{th}) hyperparameter, and *(p)* is the number of values for each hyperparameter, the grid comprises all possible combinations of these values. While computationally expensive, grid search offers a thorough exploration, especially when prior knowledge about the ideal hyperparameter range is limited.

Fine-tuning strategies ensure model is robust, accurate and optimized for predicting graduate employment outcomes.

3.7. Evaluation Metrices

Rigorous model evaluation is essential for determining the efficacy of our predictive model. We'll use a combination of metrics to provide a nuanced understanding of its strengths and weaknesses:

- ***Accuracy:*** The overall proportion of correct predictions (both true positives and true negatives) offers a baseline performance measure (32). Mathematically, it is represented as:

$$[\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}]$$

- ***Recall:*** Crucial for ensuring the model correctly identifies positive cases (i.e., employed graduates) (33). High recall helps us minimize the number of graduates who are likely to find employment but are incorrectly classified as unlikely to be employed. The formula for Recall is:

$$[\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}]$$

- ***F1 Score:*** Balances precision and recall, particularly valuable if there's an uneven distribution of employment outcomes in the data. A high F1-score indicates the model successfully identifies employed graduates while minimizing false positives (34). The F1 Score is defined as: $[\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}]$

These metrics will assess our model's strengths and weaknesses, guiding further refinement for improved performance. This standard evaluation approach ensures rigor and aligns with best practices in the field.

4. Findings and Outcomes

This exploration into the correlation between academic performance, professional skills demand, alumni networks' influence, and the integration of homomorphic encryption (HE) unveils pivotal insights for predicting college graduates' employment outcomes. Analysis of the US Bachelor's Graduates Dataset (35), which contains information about Bachelor's degree graduates from various universities in the USA and their placement status, by applying Logistic Regression Classifier (36), confirms the importance of academic excellence, technical proficiency, and robust alumni connections in enhancing employability. Correlation between factors effecting employment rate is shown in (Fig 3).

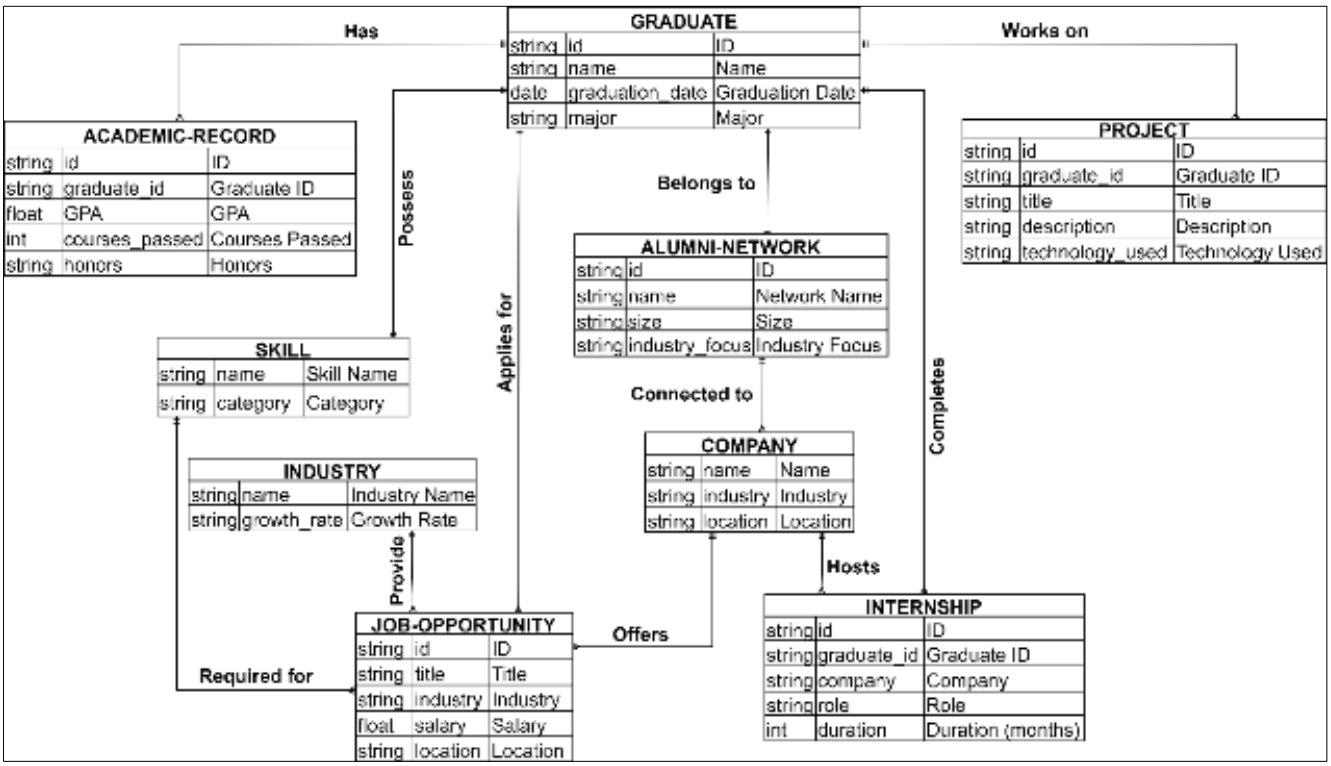

Figure 3 ER diagram of correlation between entities involved in graduates' employment

Importantly, the use of HE validates our framework's emphasis on secure data handling. It allows for in-depth analysis while safeguarding student privacy. These findings advocate for a multifaceted approach in education and career planning, emphasizing curricula that cater to market demands while prioritizing data privacy and security in digital age.

4.1. Academic Performance and Employment Rate Correlation

Academic performance has a profound impact on a graduate's entry into the workforce. A strong academic record signals intellectual capability and work ethic, making students highly attractive to employers (Fig 5a). Our big data analysis confirms a positive correlation between academic standing (e.g., GPA) and employment rates (Fig 4).

This suggests that high achievers gain a competitive edge in the job market. Additionally, a major's prestige within a specific industry plays a crucial role. Graduates from well-regarded programs in high-demand fields often have a wider array of job opportunities. Therefore, students should carefully consider both their academic performance and the reputation of their chosen major when making career decisions.

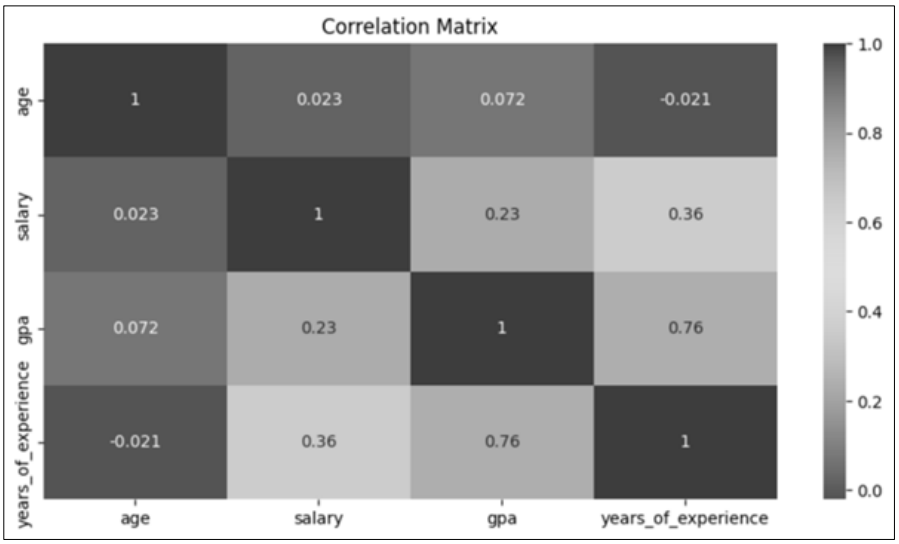

Figure 4 Correlation matrix for graduates' employment

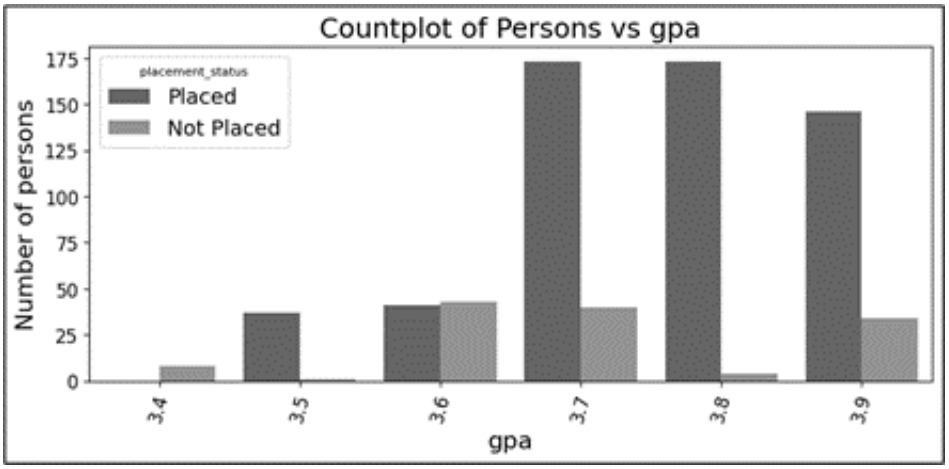

Figure 5a Employment rates as per gpa

Figure 5b Employment rates as per years of experience

Importantly, our analysis underscores that timely practical experience gained during a student's academic journey is vital for boosting employability (Fig 5b, 5c). Internships, collaborative projects, and research initiatives translate

theoretical knowledge into marketable skills. This aligns with our framework's emphasis on building a workforce equipped to apply their knowledge in real-world settings.

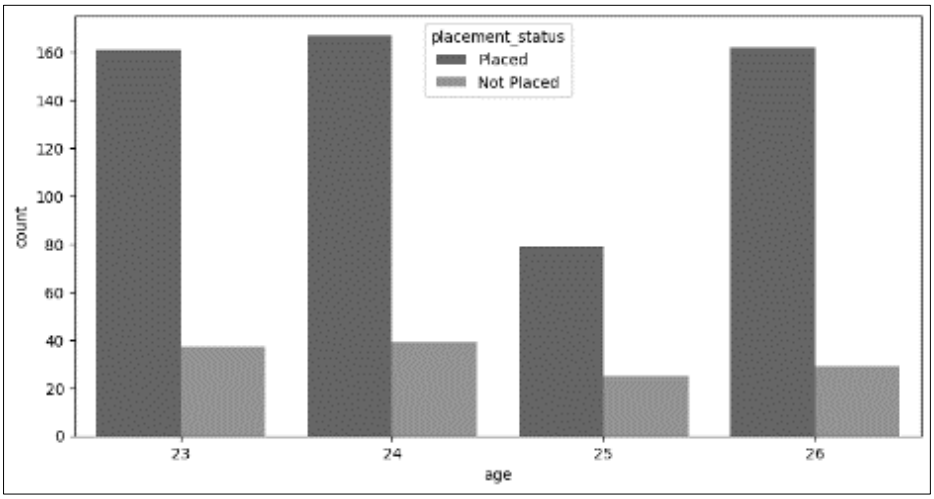

Figure 5c Employment rates as per age

In conclusion, a combination of strong academics, strategic major selection, and hands-on experience lays a solid foundation for successful career placements. Educational institutions hold the key to fostering this environment, not only emphasizing academic excellence but also actively encouraging practical learning activities that prepare students for their future careers.

4.2. Employment Market Demand Trends

Navigating the complexities of job market is crucial for students to make informed decisions about their majors. Extensive data analysis reveals key trends shaping employment opportunities and influencing future career paths.

4.2.1. High Demand for Graduates in Popular Industries

Industries like the internet, artificial intelligence (AI), and financial technology (FinTech) are rapidly expanding, creating significant demand for graduates with specialized knowledge. These fields offer promising career trajectories marked by competitive salaries and rapid growth. To meet this demand, educational institutions must be adaptive, ensuring that their curricula prepare students for the specific skills required by these booming sectors.

4.2.2. Increasing Employment Demand in Technical Majors

Technological advancements continue to drive a relentless increase in demand for technical proficiency.

Majors centered around data analysis, machine learning, and AI are cornerstones of the innovation economy. The job market consistently favors skilled graduates in these domains, emphasizing the crucial need for curricula that remain agile in response to technological progress (Fig 6).

Fluctuations in economic development and industrial restructuring bring both challenges and opportunities for graduates. While some traditional industries may see a decline in job prospects, others are rapidly emerging. Students must adopt a forward-looking approach, carefully considering fields with burgeoning demand and aligning their major selection with both their strengths and market realities. Understanding these trends empowers students to craft a strategic career roadmap. By proactively cultivating in-demand skills and seeking practical experiences, such as internships or projects, students gain a significant competitive edge. This initiative fosters resilience and adaptability – essential qualities for success in a rapidly evolving job market. Integrating these insights into educational planning and individual strategies is critical for ensuring alignment between student aspirations and market needs. This will cultivate a skilled workforce ready to navigate and contribute to the future economy, a key focus of our framework.

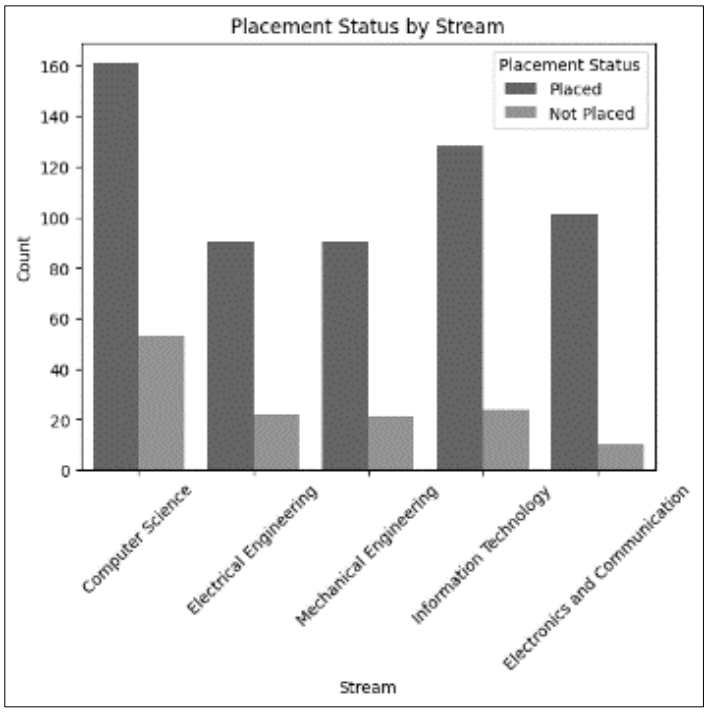

Figure 6 Employment rates of different majors

4.3. Impact of Alumni Networks

Alumni networks play a significant role in shaping the career trajectories of graduates. Our in-depth analysis underscores their multifaceted benefits:

4.3.1. Enhanced Network Scale and Outreach

Expansive alumni networks offer a wider range of career development opportunities. Robust networks create a diverse platform for communication and resource sharing, crucial for navigating the job market. Universities that actively foster alumni engagement through events and digital platforms enhance connectivity, which our framework recognizes as pivotal for student success.

4.3.2. Quality of Connections

The value of alumni networks lies not just in their size, but the depth of connections fostered. Mentorship opportunities, career advice, and potential referrals from experienced alumni are invaluable (37). Proactive relationship-building within these networks is key for students seeking to leverage this collective knowledge.

4.3.3. Strategic Industry and Geographic Alignment

Alumni networks with strong ties to specific industries or geographic regions offer targeted employment opportunities. Students should strategically focus their networking efforts to align with their career aspirations, using alumni insights to guide their decisions. For example, a student interested in the technology sector in Silicon Valley would benefit greatly from connecting with alumni working in that region and industry (15).

In essence, alumni networks offer far more than job leads – they provide a platform for lifelong professional development. Universities play a vital role in nurturing these networks, empowering students to transition successfully from academia to the workforce.

4.4. Professional Skills Demand Analysis

Our analysis of the evolving job market reveals a critical need for specific professional skills, prompting a strategic approach to education and skill development (Fig 7). Key findings from our Big Data exploration include:

4.4.1. Data Analysis and Statistical Skills

The Big Data era fuels a high demand for data analysis and statistical expertise. Mastery of these skills offers a competitive advantage across various sectors, empowering graduates to make data-driven decisions.

4.4.2. Machine Learning and AI Expertise

The rapid expansion of artificial intelligence (AI) increases the need for professionals skilled in machine learning and deep learning techniques. Students who invest in projects, courses, and AI-related training significantly boost their marketability and potential for innovation.

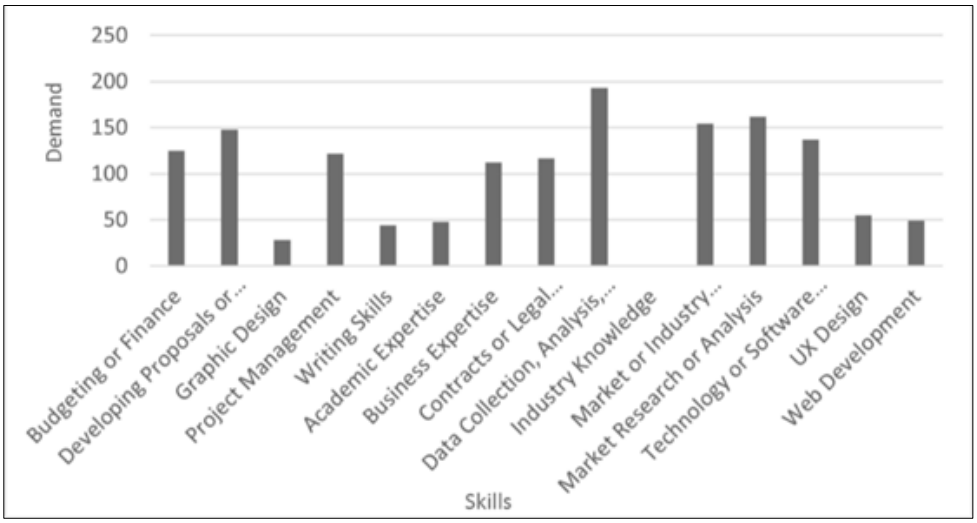

Figure 7 Demand of different skills

4.4.3. Programming Proficiency

Coding skills have become essential across industries, opening opportunities in software development, data science, AI, and more. Dedicated coding practice and targeted learning paths are crucial for students to master this versatile skill.

This analysis underscores the necessity for educational institutions to adapt their curricula to prioritize these in-demand skills. Collaboration with industry leaders is vital to provide practical learning experiences that prepare students not just for immediate employment, but position them as competitive leaders in the future job market. This emphasis on skills aligns with our framework's focus on adaptability and workforce development.

4.5. Impact of Applying Homomorphic Encryption in Employment Prediction

The integration of homomorphic encryption (HE) into employment prediction analysis signals a major shift towards prioritizing data security without sacrificing analytical insights. This technology enables computation on encrypted data, safeguarding sensitive student information while addressing the critical challenge of data breaches. While HE offers unparalleled security benefits, our analysis confirms increased computational overhead (Table 1).

Table 2 Comparison of classification report with and without applying HE

Scheme	Precision 0, 1	Recall 0, 1	F1 Score 0, 1	Support 0, 1	Accuracy	Macro Avg	Weigh Avg
Without HE	0.50, 0.92	0.65, 0.90	0.57, 0.91	30, 110	0.86	0.71	0.83
With HE	0.48, 0.90	0.63, 0.88	0.55, 0.89	30, 110	0.84	0.69	0.82

This highlights a crucial trade-off between operational efficiency and data protection. At first glance, the precision of employment predictions may appear compromised by the complexity of HE. However, the secure environment it creates allows for the inclusion of richer, more detailed datasets, potentially maintaining the accuracy and depth of insights.

Beyond technical considerations, HE aligns with our framework's emphasis on ethical data use. By minimizing the risk of unauthorized access to sensitive student information, HE supports responsible prediction models that mitigate discriminatory outcomes. A comprehensive assessment of HE's impact requires detailed comparison of key metrics before and after its implementation (Table 2).

Table 3 Comparison of HE application with traditional approaches

Metric	Without HE	With HE
Computational Time	Shorter	Longer due to encryption overhead
Data Security Level	Lower	High, as data remains encrypted
Accuracy of Predictions	Higher	Slightly Lower, enabled by secure data use
Scalability	High, limited by data privacy concerns	Lower, due to increased computational demands
Complexity of Data Operations	Lower	Higher, due to encryption processes
User Accessibility	High, with potential privacy risks	Restricted, ensures data privacy
Maintenance Costs	Lower	Higher, due to encryption management

This analysis should consider computational time, data security levels, prediction accuracy, scalability, complexity, user accessibility, and management costs.

While HE introduces challenges, it represents a significant advancement in securing sensitive data within this critical domain. Careful deployment of this technology mandates careful balancing of data privacy and analytical effectiveness. This highlights the urgent need to continue developing optimized HE algorithms that reduce computational overhead without sacrificing data protection or model accuracy. By addressing these challenges, HE's full potential for secure, ethical, and accurate employment prediction models can be realized.

5. Conclusion and Future Study

This study provides valuable insights into the multifaceted factors influencing employment outcomes for college graduates. Our analysis underscores the enduring importance of academic excellence as a predictor of employment success, reinforcing the value of a strong academic foundation. However, the job market is evolving, and our findings highlight the critical need for specialized skills in data analysis, machine learning, and programming. Graduates who proactively develop these highly sought-after skills alongside their academic pursuits will gain a significant competitive advantage. Additionally, the influence of robust alumni networks in providing mentorship, insights, and opportunities emphasizes the importance of strategic networking for career success. Importantly, the successful integration of homomorphic encryption (HE) signals a major advancement in secure and ethical employment prediction. While HE introduces computational challenges, its ability to safeguard sensitive student information aligns strongly with our framework's emphasis on data privacy and responsible analytics.

Our research opens up numerous avenues for further exploration to enhance these predictive models and deepen our understanding of factors leading to employment success. Continued refinement of HE algorithms, alongside the exploration of alternative encryption methods, could maximize efficiency while upholding data security standards. Incorporating a broader range of variables, such as socioeconomic factors and emerging industry trends, and conducting longitudinal studies to track graduates' career paths would further enhance the model's predictive power. Close collaboration with industry partners would ensure better alignment between educational outcomes and real-world needs, fostering a workforce equipped with relevant, in-demand skills. Finally, investigating the model's scalability across diverse educational contexts and geographies could significantly broaden its impact as a tool for bridging the gap between academia and industry on a global scale.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Mezhoudi N, Alghamdi R, Aljunaid R, Krichna G, Düştegör D. Employability prediction: a survey of current approaches, research challenges and applications. Journal of Ambient Intelligence and Humanized Computing. 2023;14(3):1489-505.

[2] Raman R, Pramod D. The role of predictive analytics to explain the employability of management graduates. Benchmarking: An International Journal. 2022;29(8):2378-96.

[3] Huang M-H, Rust RT. Artificial intelligence in service. Journal of service research. 2018;21(2):155-72.

[4] Fang H, Qian Q. Privacy preserving machine learning with homomorphic encryption and federated learning. Future Internet. 2021;13(4):94.

[5] Voigt P, Von dem Bussche A. The eu general data protection regulation (gdpr). A Practical Guide, 1st Ed, Cham: Springer International Publishing. 2017;10(3152676):10-5555.

[6] Goldman E. An introduction to the california consumer privacy act (ccpa). Santa Clara Univ Legal Studies Research Paper. 2020.

[7] Beaudin K. College and university data breaches: Regulating higher education cybersecurity under state and federal law. JC & UL. 2015;41:657.

[8] Iyengar SS, Wells RE, Schwartz B. Doing better but feeling worse: Looking for the "best" job undermines satisfaction. Psychological Science. 2006;17(2):143-50.

[9] Sandvig JC, Tyran CK, Ross SC. Determinants of graduating mis students starting salary in boom and bust job markets. Communications of the Association for Information Systems. 2005;16(1):29.

[10] Patacsil FF, Tablatin CLS. Exploring the importance of soft and hard skills as perceived by IT internship students and industry: A gap analysis. Journal of Technology and Science education. 2017;7(3):347-68.

[11] Goldfarb A, Taska B, Teodoridis F. Could machine learning be a general purpose technology? a comparison of emerging technologies using data from online job postings. Research Policy. 2023;52(1):104653.

[12] Pandya B, Patterson L, Ruhi U. The readiness of workforce for the world of work in 2030: perceptions of university students. International Journal of Business Performance Management. 2022;23(1-2):54-75.

[13] Persaud A. Key competencies for big data analytics professions: A multimethod study. Information Technology & People. 2021;34(1):178-203.

[14] Brink R. A Multiple Case Design for the Investigation of Information Management Processes for Work-Integrated Learning. International journal of work-integrated learning. 2018;19(3):223-35.

[15] English P, de Villiers Scheepers MJ, Fleischman D, Burgess J, Crimmins G. Developing professional networks: the missing link to graduate employability. Education+ Training. 2021;63(4):647-61.

[16] Malhotra R, Massoudi M, Jindal R. An alumni-based collaborative model to strengthen academia and industry partnership: The current challenges and strengths. Education and Information Technologies. 2023;28(2):2263-89.

[17] Schneider J, Harvan M, Obermeier S, Locher T, Pignolet Y-a. Machine learning based on homomorphic encryption. Google Patents; 2023.

[18] Akavia A, Leibovich M, Resheff YS, Ron R, Shahar M, Vald M. Privacy-preserving decision trees training and prediction. ACM Transactions on Privacy and Security. 2022;25(3):1-30.

[19] Iezzi M, editor Practical privacy-preserving data science with homomorphic encryption: an overview. 2020 IEEE International Conference on Big Data (Big Data); 2020: IEEE.

[20] Patel K. Ethical reflections on data-centric AI: balancing benefits and risks. International Journal of Artificial Intelligence Research and Development. 2024;2(1):1-17.

[21] Greenwood PE, Nikulin MS. A guide to chi-squared testing: John Wiley & Sons; 1996.

[22] Kraskov A, Stögbauer H, Grassberger P. Estimating mutual information. Physical review E. 2004;69(6):066138.

[23] Schmidt M, Fung G, Rosales R. Optimization methods for l1-regularization. University of British Columbia, Technical Report TR-2009-19. 2009.

[24] Juszczak P, Tax D, Duin RP, editors. Feature scaling in support vector data description. Proc asci; 2002: Citeseer.

[25] Kim M, Song Y, Wang S, Xia Y, Jiang X. Secure logistic regression based on homomorphic encryption: Design and evaluation. JMIR medical informatics. 2018;6(2):e8805.

[26] González S, García S, Del Ser J, Rokach L, Herrera F. A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. Information Fusion. 2020;64:205-37.

[27] Mayr A, Binder H, Gefeller O, Schmid M. The evolution of boosting algorithms. Methods of information in medicine. 2014;53(06):419-27.

[28] Pavlyshenko B, editor Using stacking approaches for machine learning models. 2018 IEEE second international conference on data stream mining & processing (DSMP); 2018: IEEE.

[29] Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S, editors. The'K'in K-fold Cross Validation. ESANN; 2012.

[30] Passos D, Mishra P. A tutorial on automatic hyperparameter tuning of deep spectral modelling for regression and classification tasks. Chemometrics and Intelligent Laboratory Systems. 2022;223:104520.

[31] Liashchynskyi P, Liashchynskyi P. Grid search, random search, genetic algorithm: a big comparison for NAS. arXiv preprint arXiv:191206059. 2019.

[32] De Sa C, Leszczynski M, Zhang J, Marzoev A, Aberger CR, Olukotun K, Ré C. High-accuracy low-precision training. arXiv preprint arXiv:180303383. 2018.

[33] Davis J, Goadrich M, editors. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning; 2006.

[34] Lipton ZC, Elkan C, Narayanaswamy B. Thresholding classifiers to maximize F1 score. arXiv preprint arXiv:14021892. 2014.

[35] (NCSES) NCfSaES. National Survey of College Graduates: 2021. 2022. Report No.: NSF 23-306.

[36] Shaukat A, Saif U, editors. NLP based Model for Classification of Complaints: Autonomous and Intelligent System. 2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2); 2022: IEEE.

[37] Singer TS, Hughey AW. The role of the alumni association in student life. New directions for student services. 2002;2002(100):51-68.