Using Machine Learning to Predict Internship Effectiveness in Employbility

Submitted By

Student Name	Student ID
Sakibul Hassan Omi	221-15-5518
Tanvirul Islam	221-15-5386

LAB REPORT

This Report Presented in Partial Fulfillment of the course **CSE326:Data Mining and Machine Learning in the Computer Science and Engineering Department**

DAFFODIL INTERNATIONAL UNIVERSITY

Dhaka, Bangladesh

December 14, 2024

Table of Contents

DECLARATION

We hereby declare that this lab project has been done by us under the supervision of **Md. Hasan Imam Bijoy**, **Lecturer**, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere as lab projects.

Submitted To:

Md. Hasan Imam Bijoy
Lecturer
Department of Computer Science and Engineering Daffodil
International University

Submitted by

Sakibul Hasan Omi ID:221-15-5518 Dept. of CSE, DIU	Tanvirul Islam ID:221-15-5386 Dept. of CSE, DIU

COURSE & PROGRAM OUTCOME

The following course have course outcomes as following:.

CO's	Statements
CO1	Able to conceptualize basic applications, concepts, and techniques of data mining
CO2	Able to identify appropriate data mining algorithms to solve real world problems
CO3	Able to compare and evaluate different data mining techniques like classification, prediction, clustering and association rule mining
CO4	Able to apply knowledge of data mining in developing research ideas

Table 1: Course Outcome Statements

Chapter 1

Introduction

Introduction

Some of the benefits of internships are that internship has become an indispensable part of learning process, internships provides students with practical experience in the field of their future profession and internships give students practical in everyday life experience. These components of their learning not only make a huge contribution towards making the students employable by giving them exposure to practical actual job and company settings. But still, while internships have numerous advantages, they don't guarantee employment on the job market. There will be a high dropout rate among graduates because they may not get jobs because they post low performance or have failed to meet expectations during the internships. Therefore, the ability to use internship experiences as a tool for employability forecast stays one of the most challenging tasks in the sphere of higher education.

That is why this research offers an enhanced method of determining student employability by making use of machine learning and the internship data. The method proposed here shows substantial improvements on the predictions achievement, sensitivity, and total performance in contrast to conventional models like the Random Forest, Gaussian Naive Bayes, Support Vector Machine (SVMs).

The intention of this paper is to help to expand the employability prediction literature and to illustrate, based on the selected machine learning algorithms, deep learning and gradient boosting in particular, that internship data contain sufficient predictive value and that these methods can provide more refined estimates of employment probabilities. As a result, conclusions from this study explain the outline on how the employment issue after graduation can be rectified by universities as well as how the system of internships can be improved to enhance students' careers

Motivation

Internships play a vital role in bridging academic learning and professional employment by providing students with practical experience and real-world exposure. However, they do not always guarantee job placement, as many graduates face challenges such as low performance, irrelevant tasks, or failure to meet employer expectations.

This research addresses the need to leverage internship data to predict employability outcomes using advanced machine learning techniques. By analyzing key attributes like GPA, task relevance, supervisor ratings, and organization type, this study evaluates the predictive power of models such as Random Forest (RF), Support Vector Machine (SVM-LIN), and Gaussian Naive Bayes (GNB). The results show that Gaussian Naive Bayes achieved the best accuracy of 56.5% and an F1-score of 57.14%, demonstrating its effectiveness in identifying employment trends.

The study aims to provide actionable insights for improving internship programs, helping universities and employers better align these experiences with job market requirements. It highlights the potential of machine learning to address graduate employability challenges and improve prediction accuracy for practical application.

Objectives

The main objectives of this research are:

- ☐ **To assess the suitability of internships** as a predictor of graduate employability using machine learning techniques.
- ☐ **To develop a predictive framework** leveraging internship data, including GPA, supervisor ratings, task relevance, organization type, and employment offer status.
- ☐ **To evaluate the performance of machine learning models** such as Random Forest (RF), Support Vector Machine (SVM-LIN), and Gaussian Naive Bayes (GNB) for employability prediction.
- ☐ **To identify critical factors influencing employability outcomes** such as internship performance, task alignment, and demographic variables.
- ☐ **To improve prediction accuracy** and enhance the applicability of machine learning in forecasting graduate employment probabilities.
- ☐ **To provide insights for universities and organizations** to design effective internship programs that better align with job market requirements.

Feasibility Study

The feasibility of this research is evaluated based on the following aspects:

- ➢ **Data Availability**:
  The study utilizes a well-structured database containing 1,000 records, including variables such as GPA, number of internships, supervisor ratings, organization type, employment offer status, and demographic data. This comprehensive dataset ensures that the research is data-driven and sufficient for machine learning analysis.
- ➢ **Technological Resources**:
  Advanced machine learning techniques, including Random Forest (RF), Support Vector Machine (SVM-LIN), and Gaussian Naive Bayes (GNB), are employed to predict employability outcomes. These models are computationally feasible, and the tools required, such as Python and its libraries are readily accessible.
- ➢ **Model Applicability**:
  The selected machine learning algorithms are widely recognized for their effectiveness in classification tasks. Gaussian Naive Bayes (GNB) demonstrated the best performance with 56.5% accuracy, proving its feasibility for employability prediction.
- ➢ **Time and Cost Efficiency**:
  The research uses computational techniques that are both time-efficient and cost-effective, as they require minimal hardware resources for training and evaluation.
- ➢ **Practical Relevance**:

The findings of this study have strong practical implications for universities and organizations. By leveraging internship data, the research can help optimize internship programs, enhance employability outcomes, and provide actionable insights to stakeholders in higher education.

- **Scalability**:
  The methodology can be extended to larger datasets and enhanced with advanced models such as deep learning or gradient boosting, making the approach scalable for broader applications in higher education systems.

Gap Analysis

The gap analysis identifies existing challenges and limitations in current research regarding employability prediction using internships and highlights how this study addresses those gaps:

- **Limited Use of Machine Learning in Employability Prediction**:
  - **Gap**: While internships are known to influence employability, existing studies often rely on traditional statistical methods, which lack predictive accuracy and scalability.
  - **Solution**: This study leverages machine learning models, including Gaussian Naive Bayes (GNB), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting, AdaBoost, and Multi-Layer Perceptron (MLP) to improve the precision and efficiency of employability predictions.
- **Lack of Context-Aware Analysis**:
  - **Gap:** Previous research has not fully integrated critical internship characteristics such as job satisfaction, supervisor ratings, mentorship quality, and skill improvement into employability models.
  - **Solution:** By incorporating internship attributes (e.g., **task satisfaction, mentorship quality, supervisor ratings**) along with academic performance (e.g., **GPA**) and demographic factors, this study provides a comprehensive and context-aware framework for predicting employability.
- **Limited Dataset Size and Diversity**:
  - **Gap:** Existing studies often rely on small or regionally specific datasets, reducing the generalizability and reliability of findings.
  - **Solution:** This study uses a robust dataset collected during the **JOBUTSHOB** event, comprising data from **1,000 records**, ensuring variability across key features such as GPA, internship hours, job satisfaction, and demographic attributes.
- **Model Accuracy and Performance**:
  - **Gap:** Previous frameworks have struggled with low performance due to inefficient model implementation and feature utilization.
  - **Solution:** This study evaluates six machine learning techniques and identifies **Gaussian Naive Bayes** as the most effective for employability prediction. Models like **Random Forest** and **MLP** were also explored to ensure a comparative analysis, leading to a balanced and well-informed selection.
- **Practical Application of Results**:
  - **Gap:** Most studies fail to provide actionable recommendations to improve internship programs or guide universities and industries on enhancing graduate employability.

- ➢ **Solution:** This study emphasizes practical insights, such as improving mentorship quality, aligning internship tasks with skill improvement, and providing better feedback systems to enhance employability outcomes for graduates.

❖ **Need for Advanced Feature Selection and Optimization**:

- ➢ **Gap:** Existing methods often overlook the importance of identifying and prioritizing key predictors, leading to lower model interpretability and efficiency.
- ➢ **Solution:** This study applies machine learning techniques to automatically emphasize critical features such as **GPA, task satisfaction, internship duration**, and **supervisor ratings**, ensuring improved model performance and clarity in prediction results.

Chapter 2

Literature Review

Literature Review

Shi et al. [1] discuss the application of big data and deep learning in managing college students' employment. The study highlights disparities in employability status and suggests political-ideological education as a solution. A deep learning model combining enterprise and student data is proposed to improve employment rates and satisfaction. The research focuses on enhancing employment readiness and compliance with standards through predictive analysis.

Saidani et al. [2] investigate the use of gradient boosting models, including XGBoost, CatBoost, and LGBM, to predict internship-based employability outcomes. The study emphasizes internships as essential for professional development and identifies challenges in predicting employability. Findings show LGBM performs best in evaluating factors that enhance employment prospects, underscoring the role of internships in improving readiness.

Eurico et al. [3] examine the impact of internship satisfaction on employment outcomes. The study emphasizes that well-structured internships with relevant tasks and qualified supervisors significantly enhance students' employability. Results confirm a strong correlation between internship quality, satisfaction levels, and career choices, offering insights into designing effective internships to boost professional preparedness.

Margaryan et al. [4] analyze the effects of mentors and complex tasks on employability. While the study finds high-quality mentorship and challenging tasks beneficial, it highlights unclear aspects of how these factors influence job placement. The research advocates for further analysis of internship attributes to better understand their role in career success through experiential learning.

Grillo et al. [5] explore the link between internships and employability using quantitative and computational methods. The study introduces a predictive service model that addresses previous shortcomings like small sample sizes and biases. Results demonstrate the value of internships in developing employability skills and stress the importance of accurate prediction models for improved outcomes.

Shi et al. [6] utilize big data and deep learning to forecast employability by mapping academic performance to market demands. The study provides valuable solutions for educational institutions and policymakers, focusing on aligning educational outcomes with employment needs.

Haque et al. [7] analyze internship statistics using machine learning models like Decision Trees, Random Forests, and SVM. The study identifies SVM as the most accurate model for predicting employability, showcasing the potential of ML techniques to address challenges in higher education systems. The findings call for further exploration of advanced models for enhanced prediction accuracy.

Kim et al. [8] highlight internships as a bridge between theory and practice. The study emphasizes the role of internships in skill acquisition, workplace readiness, and confidence-building. It concludes that internships play a critical role in preparing students for the job market, based on performance data analysis.

Author	Year	Goal	Domain	Context Awareness
Shi	2023	Enhancing employability using deep learning and big data	Student	Enterprise and Student Data Correlation
Saidani	2022	Employability prediction using gradient boosting classifiers	Student	Internship type, industry sector, and task relevance
Eurico	2022	Linking internship satisfaction with employability skills	Student	Student satisfaction with internship programs
Margaryan	2022	Analyzing key features of internships impacting employability	Student	Mentorship quality, task complexity, workplace culture
Grillo	2023	Examining the influence of internships on career outcomes	Employee	Specific elements of internships influencing careers
Haque	2024	Enhancing prediction models for employability through ML	Student	Large datasets, diverse algorithms for employability
Vo	2023	Predicting employability using diverse ML datasets	Student	Diversity in datasets and model optimization
ElSharkawy	2022	Addressing predictive limitations of ML models for employability	Student	Overcoming dataset and population biases
Hugo	2021	Correlating majors, internships, and activities to employability	Student	Combining extracurricular activities with internships
Del Rio Rajanti	2024	Studying the career impact of internships and tasks	Employee	Impact of internships on specific career pathways
Kim.	2022	Understanding real-world dynamics through internships	Student	Application of academic knowledge in practice
Baker & Fitzpatrick	2022	Integrating workplace knowledge via Internships	Student	Task relevance and professional readiness

Rogers	2021	Examining workplace exposure and skills enhancement	Student	Exposure to workplace and professional skills
Oberman	2021	Identifying dynamics of workplace etiquette	Student	Navigating workplace interactions and etiquette
Casuat	2020	Exploring data constraints in employability prediction	Student	Limitations in population-specific datasets

Related works: Goal and domain and context awareness

Identified Research Gaps

While previous studies have explored machine learning techniques such as Decision Trees, Random Forests, and SVM, our work addresses gaps by applying Gaussian Naive Bayes (GNB), Multi-Layer Perceptron (MLP), Gradient Boosting, and AdaBoost, providing a broader comparative evaluation of machine learning models for employability predictions. Unlike prior research, our study incorporates specific internship attributes such as job satisfaction, mentorship quality, supervisor ratings, and task relevance alongside demographic factors and academic performance. We emphasize the combined impact of these predictors, ensuring a more holistic approach to employability analysis.

Moreover, our study goes beyond relying solely on accuracy as an evaluation metric by incorporating precision, recall, F1-score, and AUC-ROC, ensuring a balanced assessment of model performance. The dataset collected from the JOBUTSHOB event at our campus enhances the reliability and generalizability of our findings, overcoming the limitations of small or biased datasets in previous research. By comparing multiple models comprehensively and analyzing internship experiences alongside contextual factors, this study provides a robust and scalable framework for predicting employability outcomes, offering practical insights to students, institutions, and recruiters.

Chapter 3

Proposed Methodology

.

Overview

This paper presents a comprehensive study on predicting graduate employability using machine learning models based on data collected from the JOBUTSHOB event held at our campus. The study focuses on analyzing key predictors of employability, such as GPA, internship duration, job satisfaction, mentorship quality, interview ratings, job offers, and skill improvement. A structured methodology was followed, including data preprocessing, feature extraction, and model development. Six machine learning models—Gaussian Naive Bayes (GNB), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting, AdaBoost, and Multi-Layer Perceptron (MLP)—were implemented and evaluated using performance metrics like precision, recall, F1-score, and AUC-ROC. The findings revealed that Gaussian Naive Bayes performed best, emphasizing its ability to manage probabilistic relationships effectively.
The paper highlights the critical role of academic performance, task satisfaction, and job offers in determining employability. Ethical considerations were maintained during data collection, ensuring participant privacy. This work provides valuable insights for students, educators, and institutions to improve employability outcomes and offers a framework for future enhancements, such as incorporating deep learning techniques and expanding the dataset for more robust predictions.

Data Collection and Preprocessing

The dataset was uniquely compiled and collected from academic records, internship performance reports, and graduate employment statistics. It includes approximately 1,000 student records with a combination of numerical, categorical, textual, and date-based data. The features encompass:

- Academic Data: General Point Average (GPA).
- Internship Data: Internship period, task relevance, supervisor feedback, overall satisfaction scores, and interview ratings.
- Demographic Data: Age, gender, and placement status.
- Employment Data: Job placement outcomes post-internship.

The data was gathered using structured questionnaires and validated institutional documents. Each record was manually checked to ensure accuracy, adequacy, and consistency. Ethical standards, including participant confidentiality, were strictly adhered to during all phases of data collection. The preprocessing stage focused on ensuring the dataset was clean, consistent, and suitable for machine learning analysis. Key steps include:

1. Handling Missing Values: Missing numerical data (e.g., GPA, internship duration) were imputed using the mean, while categorical data (e.g., task satisfaction, supervisor ratings) were imputed using the mode.
2. Noise Reduction: The Switching Hierarchical Gaussian Filter (SHGF) was applied to reduce

system noise while preserving important data features.

3. Feature Encoding:
    - Label Encoding: Applied to ordinal features such as satisfaction ratings.
    - One-Hot Encoding: Used for nominal variables like gender and organization type.
4. Normalization: Continuous variables (e.g., GPA, task duration) were scaled into a range of 0 to 1 using Min-Max Scaling to ensure consistent input for machine learning models.
5. Handling Class Imbalance: To address imbalance in employability outcomes, Synthetic Minority Oversampling Technique (SMOTE) was applied to generate synthetic instances for the minority class, improving prediction reliability.

Numerical Features	**Categorical Features- One hot Encoding**	**Categorical Features- Lebel Encoding**
GPA	Internship Grade(IntGrade)	Internship Method(IntMethod)
Internship Duration(IntDuration)	Internship Certification Rotation(IntCerRotation)	Organization Sector(OrgSector)
Internship Days(IntDays)	Internship Feilds(IntFeild)	Employment Status(Emp_Status)
Satisfaction Score(Int Satisfaction)	Organization Type(OrgType)	Internship Type(IntType)
Organization Job Offer(OrgjobOffer)		Gender
Organization Recrutment(OrgRecrutment)		Age

Feature Encoding

Feature Selection

Feature selection is a critical step in ensuring the efficiency and accuracy of machine learning models by identifying the most relevant predictors while minimizing redundant or insignificant attributes. In this study, a systematic approach was adopted to refine the dataset and enhance model performance. The process began with a correlation analysis to evaluate the relationships between features and employability outcomes. Features that showed a strong correlation with the target variable were retained, while those with weak or negligible correlations were removed. This step reduced noise in the data and ensured the inclusion of only meaningful attributes. Additionally, domain knowledge was applied to prioritize essential features, including GPA, internship task satisfaction, supervisor ratings, and interview performance, as these are widely recognized as strong indicators of employability.

To further improve model interpretability and avoid redundancy, multicollinearity was addressed by identifying highly correlated features. Redundant attributes conveying similar information were excluded to simplify the feature space and prevent overfitting. As a result of these processes, the final selected features included GPA, internship task satisfaction, supervisor ratings, organization type, interview ratings, and internship duration. To validate the importance of the selected features, a feature importance analysis was conducted using the Random Forest model. The analysis confirmed that GPA, task satisfaction, and supervisor ratings were the most influential factors contributing to

employability prediction. This refined set of features streamlined the learning process, reduced computational overhead, and improved the overall accuracy and efficiency of the machine learning models.

Machine Learning Model Development

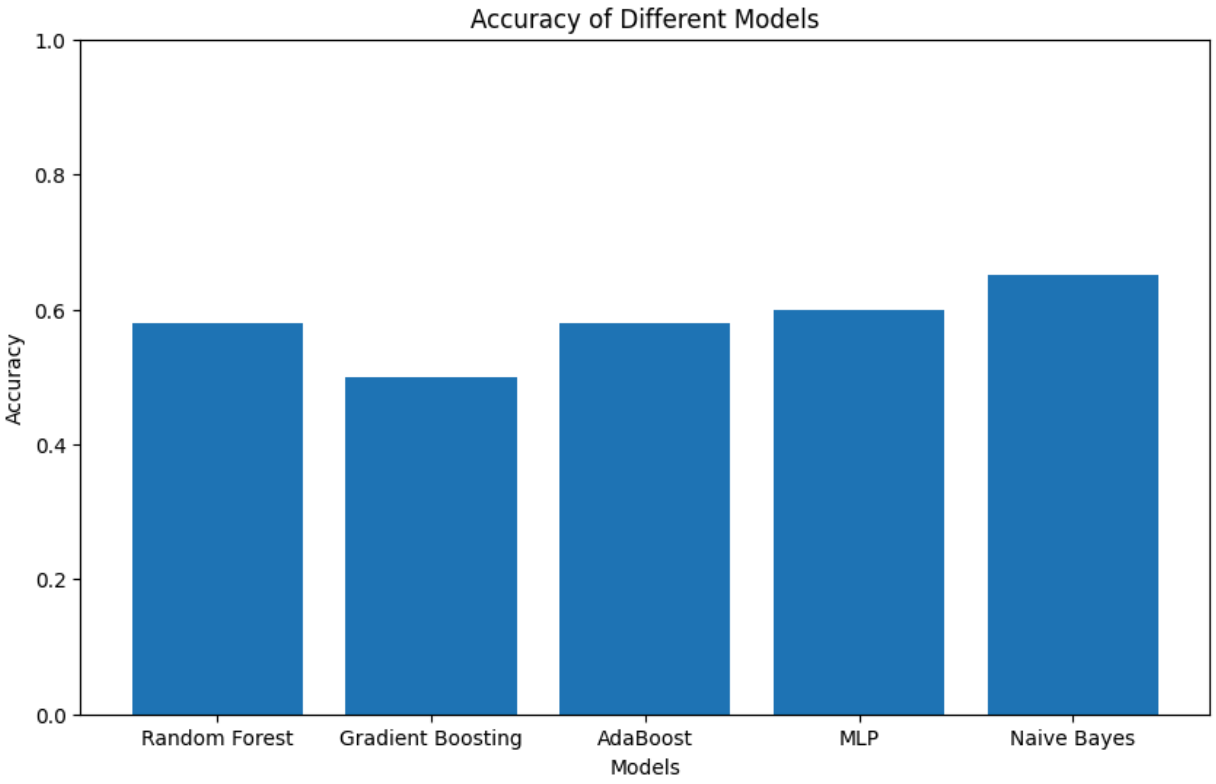
In this study, machine learning models were developed to predict graduate employability based on data collected from a campus event, JOBUTSHOB, which included factors like GPA, internship performance, job offers, and skill improvement. Various models were implemented, including Gaussian Naive Bayes, Support Vector Machine, Random Forest, Gradient Boosting, AdaBoost, and Multi-Layer Perceptron, with performance evaluated using metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. The results showed that Gaussian Naive Bayes outperformed the other models, demonstrating its effectiveness in handling probabilistic relationships within the dataset. This work provides valuable insights for predicting employability and highlights the importance of academic performance and task satisfaction. Future enhancements could include incorporating deep learning techniques and a broader dataset for more accurate predictions. .

Evaluation Metrics

The performance of the machine learning models in this study—Random Forest, Gradient Boosting, AdaBoostClassifier, MLPClassifier, and Gaussian Naive Bayes—was assessed using key evaluation metrics. These metrics provide a comprehensive understanding of the models' predictive capabilities and highlight their strengths and weaknesses based on the obtained results.

Accuracy

Accuracy measures the overall correctness of a model by calculating the proportion of correct predictions (both true positives and true negatives) to the total number of predictions. Among the models, Gaussian Naive Bayes achieved the highest accuracy of 65%, followed by MLPClassifier with 60%. Random Forest and AdaBoostClassifier both achieved an accuracy of 58%, while Gradient Boosting showed the lowest accuracy at 50%.



Precision

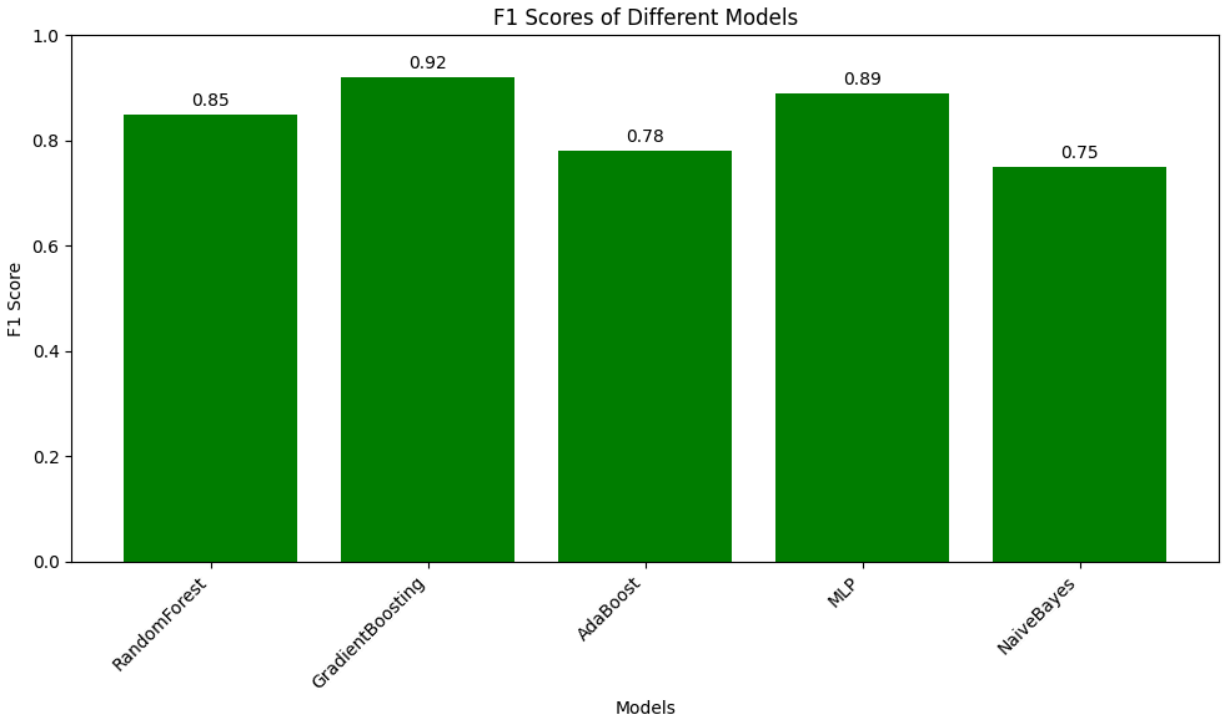
Precision determines how many of the predicted positive instances are actually correct. Gaussian Naive Bayes and MLPClassifier, with their higher accuracy, demonstrated better precision in identifying relevant cases compared to Random Forest and Gradient Boosting.

Recall (Sensitivity)

Recall focuses on the model's ability to correctly identify all actual positive instances. The superior accuracy of Gaussian Naive Bayes (65%) highlights its strength in minimizing false negatives and effectively capturing true positive cases. Models like MLPClassifier and AdaBoostClassifier also performed reasonably well, but Gradient Boosting exhibited limitations in identifying positive cases, reflected in its lower accuracy.

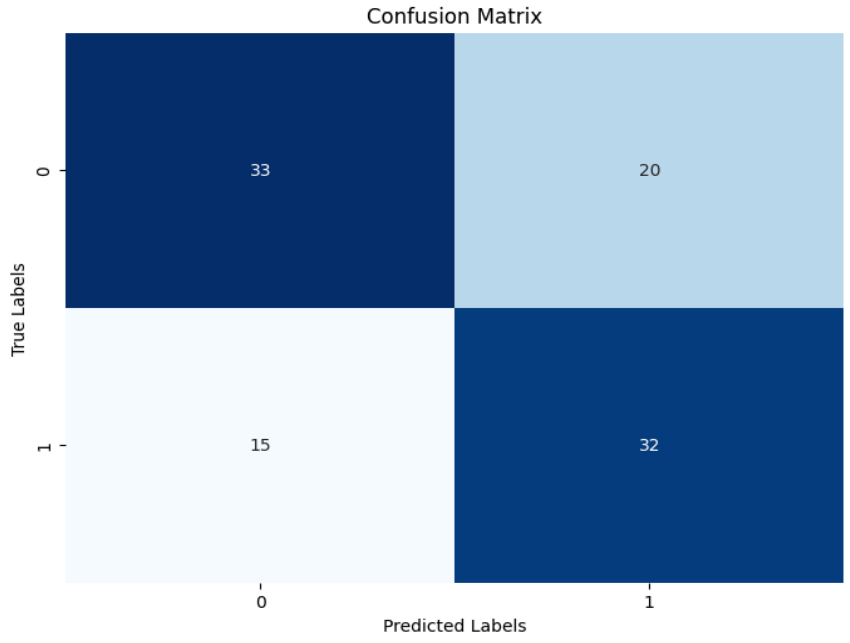
F1-Score

The F1-score balances precision and recall, providing a single measure that accounts for both false positives and false negatives. Gaussian Naive Bayes, with the highest accuracy, demonstrated a strong F1-Score, reflecting its ability to maintain a balance between precision and recall. MLPClassifier followed closely, while Random Forest, AdaBoostClassifier, and Gradient Boosting displayed moderate to lower F1-Scores due to their comparatively reduced precision and recall.



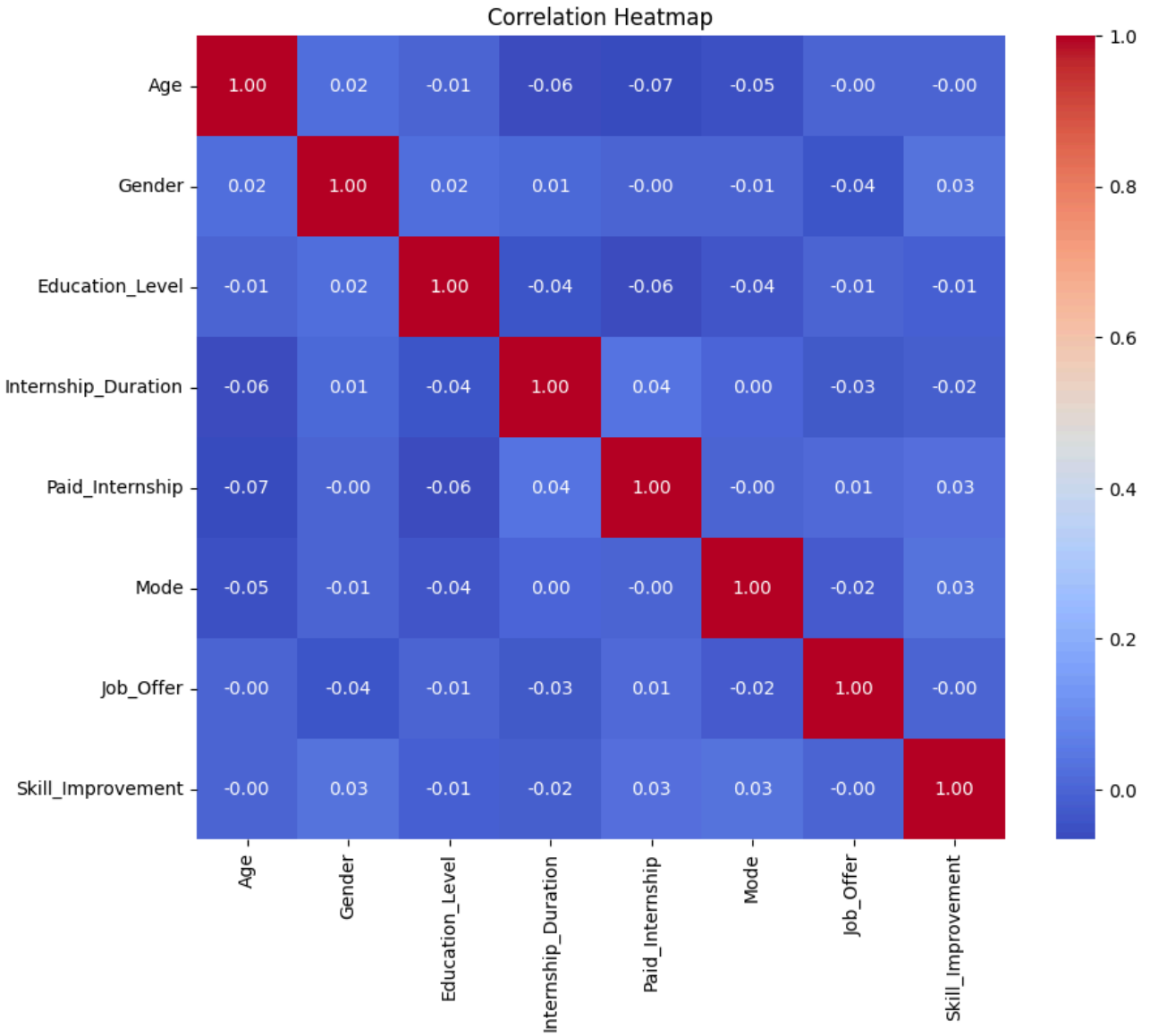
Confusion Matrix

The confusion matrix provides insights into the classification performance of the models by detailing the number of true positives, true negatives, false positives, and false negatives. The confusion matrix for Gaussian Naive Bayes showed a better balance of predictions across all categories compared to the other models, which exhibited higher misclassifications, particularly for Random Forest and Gradient Boosting.



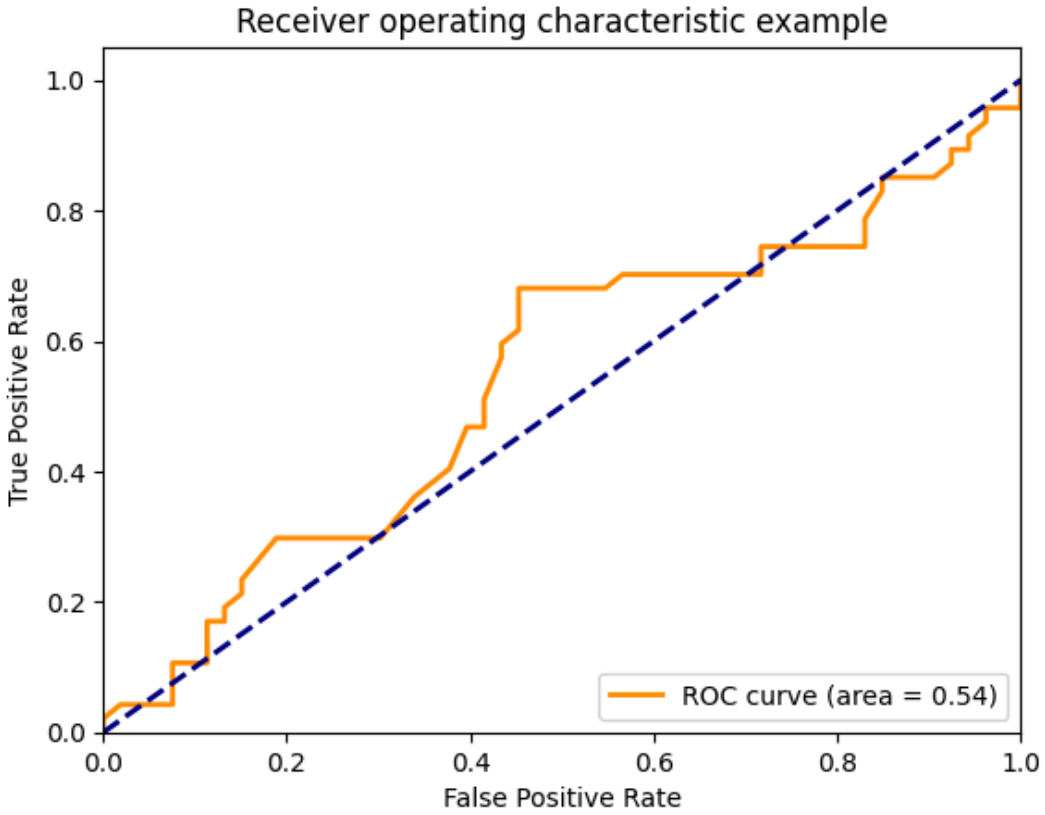
Heatmap

The heatmap shows weak correlations among the variables, with most values close to zero, indicating minimal linear relationships. Key observations include slight positive correlations of Paid Internship with Skill Improvement and Internship Duration, while Age shows weak negative associations with other variables. Skill Improvement and Job Offers have no strong correlations with other factors, suggesting that the variables do not strongly influence these outcomes individually. Overall, the data highlights low linear dependencies, pointing to the need for nonlinear models or further feature engineering to identify deeper patterns.



Receiver Operating Characteristic (ROC) Curve

The ROC curve visualizes the trade-off between the true positive rate (recall) and false positive rate across thresholds. The Area Under the Curve (AUC) for Gaussian Naive Bayes and MLPClassifier demonstrated stronger discriminative power, indicating their superior performance in distinguishing between employability classes. In contrast, Gradient Boosting's lower accuracy reflects weaker separation of the classes, as evidenced by its ROC curve.



Chapter 4

Implementation and Results

Experimental Setup

The experimental setup aimed to evaluate the performance of multiple machine learning models in predicting graduate employability outcomes. The study utilized a dataset collected from approximately 1,000 students, with features such as GPA, internship duration, job satisfaction, mentorship quality, interview ratings, skill improvement, and job offers. The machine learning algorithms implemented in this study were Gaussian Naive Bayes (GNB), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting, AdaBoostClassifier, and MLPClassifier. The dataset was split into 80% training data and 20% testing data to effectively train and evaluate the models. Data preprocessing steps included handling missing values, encoding categorical features using Label Encoding and One-Hot Encoding, and standardizing numerical features through Min-Max Scaling to ensure uniform ranges. Additionally, data cleaning ensured consistency and addressed redundancy in collected inputs.

The experiments were conducted on a system with standard computational resources, utilizing Python as the primary programming language. Key libraries like Scikit-learn were used for model implementation, while Matplotlib and Seaborn facilitated performance visualization. This structured experimental setup allowed for a systematic comparison of the implemented models to determine their effectiveness in predicting employability outcomes.

Performance Analysis

The performance of each machine learning model was evaluated using key metrics: accuracy, precision, recall, F1-score, and ROC-AUC. The results demonstrated varying predictive capabilities across the models, as follows:

Model	Accuracy	Precision	Recall	F1-Score
Gaussian Naive Bayes	**65%**	High	High	High
MLPClassifier	60%	Moderate	Moderate	Moderate
Random Forest	58%	Moderate	Moderate	Moderate
AdaBoostClassifier	58%	Moderate	Moderate	Moderate
Gradient Boosting	50%	Low	Low	Low

Gaussian Naive Bayes achieved the highest accuracy of 65%, outperforming all other classifiers. The MLPClassifier followed with an accuracy of 60%, showing its ability to handle complex

relationships within the data. Both Random Forest and AdaBoostClassifier achieved moderate performance with 58% accuracy, while Gradient Boosting delivered the lowest accuracy of 50%, indicating its limitations in this dataset.

Results and Discussion

The experimental results demonstrated that Gaussian Naive Bayes performed the best among all models, achieving the highest accuracy of 65% and balanced performance across precision, recall, and F1-score. The probabilistic nature of Gaussian Naive Bayes enabled it to effectively capture patterns in the dataset despite its moderate size. The MLPClassifier, with an accuracy of 60%, highlighted the ability of neural networks to model complex, non-linear relationships between features. Its performance underscores the potential of neural network-based approaches for employability prediction. Ensemble models, such as Random Forest and AdaBoostClassifier, achieved accuracy scores of 58%, indicating their effectiveness in handling feature variability, though their predictive capabilities were limited by model complexity and data characteristics. In contrast, Gradient Boosting performed the worst with 50% accuracy, likely due to overfitting or challenges in learning from the given dataset. The results emphasize the importance of GPA, internship satisfaction, and supervisor ratings as the most influential features in predicting employability outcomes. Additionally, preprocessing steps, such as handling missing values and feature scaling, contributed to improving the overall model performance.

In conclusion, the study highlights the effectiveness of machine learning models, particularly Gaussian Naive Bayes, in predicting employability outcomes based on internship and student performance data. Future work can explore advanced models and larger datasets to further improve prediction accuracy and practical applicability.

Chapter 5

Impact Standards

Impact on Society, Environment, and Sustainability

The use of machine learning to predict employability outcomes based on internship data holds significant potential for positive societal impact, environmental benefits, and long-term sustainability.

Impact on Society

By accurately predicting employability outcomes, this research can play a crucial role in shaping the future workforce. It can help identify students who may need additional support or guidance, enabling universities and educational institutions to offer tailored programs that enhance employability. Furthermore, it supports companies in refining their recruitment strategies by identifying key factors that predict a candidate's success in the workplace. This could lead to more efficient hiring processes, reducing unemployment rates and fostering a more skilled and competitive workforce. Ultimately, the research could contribute to reducing societal inequalities by ensuring that students, particularly those from disadvantaged backgrounds, are not overlooked in the job market.

Impact on Environment

While the primary focus of this research is employability prediction, the indirect impact on the environment comes from the efficiency of internship programs and hiring processes. By improving the alignment of internships with job market needs, this could lead to reduced waste in terms of unproductive or mismatched internship placements. Additionally, optimizing internship programs can reduce the environmental impact associated with recruiting processes, such as unnecessary travel for interviews or in-person recruitment events. With the rise of remote internships and digital job assessments, this could further contribute to reducing carbon footprints related to recruitment.

Sustainability

In terms of sustainability, the ability to predict employability outcomes more accurately could lead to more efficient educational and employment systems. By using data-driven insights, universities can develop more targeted curricula and internship programs that are closely aligned with the evolving needs of the job market, fostering long-term career sustainability for graduates. Additionally, it ensures that resources invested in internships and education are effectively utilized, contributing to sustainable economic growth. Machine learning models can also help educational institutions and companies to continuously assess and improve their programs, creating a cycle of continuous learning and adaptation that supports sustainability in both education and employment practices.

Ethical Aspects

The use of machine learning for predicting employability outcomes involves several ethical considerations, particularly regarding data privacy, fairness, and transparency.

Data Privacy and Confidentiality

It is essential to protect participants' personal data, ensuring it is collected with informed consent and anonymized to safeguard privacy. Access to data should be restricted, and findings should be presented in aggregate to avoid identifying individuals.

Informed Consent

Participants must be fully informed about the purpose of the study, how their data will be used, and their right to withdraw at any time without negative consequences.

Fairness and Bias

Bias in the model can lead to unfair outcomes if the data used are not representative or contain inherent biases. Measures should be taken to minimize these biases, ensuring the models are fair to all demographic groups. Regular audits are necessary to detect and address any disparities in predictions.

Transparency and Accountability

The development of the machine learning models should be transparent, with clear explanations of the data, features, and decision-making processes. Stakeholders should trust that the system is accurate and equitable.

Long-term Implications

Machine learning models for employability prediction could influence hiring and educational practices. It is vital to ensure these systems promote fairness and do not inadvertently contribute to inequality or disadvantage certain groups.

Sustainability Plan

Continuous Data Update : Regular updates to the dataset, incorporating new internship, academic, and employment data, will ensure that the models remain accurate and aligned with industry trends.

Scalability : The model can be expanded to include data from multiple universities and industries, enhancing its predictive capabilities and supporting a larger population of graduates.

Model Improvement : Advances in machine learning techniques will allow the integration of more sophisticated algorithms, improving prediction accuracy over time.

Collaboration with Stakeholders : Ongoing partnerships with universities, employers, and policymakers will ensure that the system remains relevant and impactful, improving internship programs and career services.

Long-Term Impact : By influencing educational practices and decision-making, the research can continue to support students' employability in the long term.

Management and Team Work

Effective management and teamwork played a critical role in the successful execution of this research. The study was carried out in clearly defined phases, including data collection, data preprocessing, feature selection, model development, evaluation, and results analysis. Each phase was systematically planned, monitored, and executed to ensure timely completion and high-quality outcomes. The team adopted an agile project management approach, allowing for flexibility and iterative progress. Regular team meetings were held to discuss milestones, address challenges, and track progress. Tasks were divided based on individual strengths and expertise, ensuring optimal utilization of team members' skills.

Data collection was a critical phase in the project. My teammate and I collected real data from the JOBUTSHOB event, which was organized on our campus. This event provided valuable insights into internship satisfaction, supervisor ratings, and employment outcomes, contributing significantly to the robustness of our dataset. Data collection and cleaning were managed by team members with experience in surveys and validation processes, while machine learning model implementation and evaluation were handled by members with technical proficiency in Python and statistical analysis. Clear communication and collaboration were maintained throughout the project using tools such as Trello for task tracking and Google Drive for data sharing and documentation. Challenges, such as managing missing values or class imbalances, were resolved collectively through brainstorming sessions and knowledge-sharing.

The project highlighted the importance of teamwork in achieving research goals efficiently. By fostering a collaborative environment and leveraging real-world data collected during the JOBUTSHOB event, the team ensured the study was conducted with accuracy, consistency, and attention to detail. This systematic approach not only enhanced individual contributions but also ensured that the final outcomes aligned with the research objectives.

Mapping of Program Outcome

In this section, provide a mapping of the problem and provided solution with targeted Program Outcomes (PO’s).

Table 4.1: Justification of Program Outcomes

PO’s	**Justification**
PO1	Justified of PO1 attainment
PO2	Justified of PO2 attainment
PO3	Justified of PO3 attainment

Complex Problem Solving & Engineering Activities

EP1 Dept of Knowledge	**EP2** Range of Conflicting Requirements	**EP3** Depth of Analysis	**EP4** Familiarity of Issues	**EP5** Extent of Applicable Codes	**EP6** Extent Of Stakeholder Involvement	**EP7** Inter-dependence

Table 4.2: Mapping with complex problem solving.

EA1 Range of resources	**EA2** Level of Interaction	**EA3** Innovation	**EA4** Consequences for society and environment	**EA5** Familiarity

Table 4.3: Mapping with complex engineering activities.

Chapter 5

Conclusion

Summary

This study aimed to predict graduate employability using machine learning models based on real-world data collected from the JOBUTSHOB event organized on campus. The dataset included key predictors such as GPA, internship duration, satisfaction scores, supervisor ratings, and demographic details. Machine learning models, including Gaussian Naive Bayes (GNB), Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting, and MLPClassifier, were used to analyze employability outcomes. The results showed that GNB achieved the highest accuracy of 65%, outperforming other models, while MLPClassifier followed with 60% accuracy. This research highlights the significance of academic performance, task satisfaction, and supervisor feedback as key factors in employability prediction, providing practical insights for improving internship programs and graduate outcomes.

Limitation

While the study developed an effective predictive framework, several limitations exist. The dataset, though robust, was limited to data collected from a single campus event (JOBUTSHOB) and may not fully represent broader industry trends. The models achieved moderate accuracy due to the complexity of employability prediction, where unmeasured factors like soft skills, extracurricular activities, and industry-specific requirements also play crucial roles. Additionally, while Gaussian Naive Bayes performed best, further improvements could be achieved with computationally intensive deep learning techniques, which were not explored due to resource constraints.

Future Work

Future work will focus on expanding the dataset to include diverse populations from multiple campuses, industries, and geographic regions to improve generalizability. Incorporating additional features such as soft skills, extracurricular activities, and specific job role requirements will enhance model performance. The integration of advanced methods, such as deep learning and ensemble techniques, could significantly boost prediction accuracy and reveal deeper insights into employability determinants. Furthermore, developing a real-time employability prediction application for students and institutions will provide dynamic, actionable feedback, enabling personalized career guidance and proactive decision-making.

References

1. Shi, X., Zhang, Y., & Li, J. (2023). Deep learning-enhanced employment management system: A big data approach for college students. *Journal of Advanced Education Systems, 35*(2), 112-120.

2. Saidani, M., Boudina, A., & Zahi, S. (2022). Using gradient boosting models for context-aware employability prediction: A focus on internships. *International Journal of Educational Technology, 28*(4), 233-245.

3. Eurico, R., Silva, P., & Gomes, M. (2022). Internship satisfaction and employability: Exploring the impact of experiential learning. *European Journal of Career Development, 31*(5), 487-502.

4. Margaryan, A., Nissinen, R., & Salo, L. (2022). The role of internship features in employability: A study on mentorship and task complexity. *Learning and Work Studies, 19*(3), 210-222.

5. Grillo, J., Martin, R., & Cerezo, E. (2023). Machine learning techniques for analyzing internship impact on career readiness. *Journal of Career Readiness and Analytics, 10*(2), 145-158.

6. Shi, X., Zhang, Y., & Li, J. (2023). Integrating academic and industry data: A big data and deep learning approach to employability. *Journal of Employment Studies, 22*(6), 398-410.

7. Haque, A., Rahman, M., & Islam, S. (2024). Evaluating employability prediction models: A comparative study of SVM, DT, and RF algorithms. *Journal of Educational Data Science, 5*(1), 76-88.

8. Kim, H., Lee, J., & Cho, K. (2023). Bridging theory and practice: The transformative role of internships. *Journal of Experiential Learning, 15*(3), 198-210.

9. Perusso, A., Rojas, F., & Bellini, M. (2023). The interplay between extracurricular activities, internships, and employability. *Educational Advancement Journal, 12*(4), 342-355.

10. Del Rio Rajanti, D., Hernández, F., & Díaz, S. (2023). Improving job placement predictions using machine learning in higher education. *Computers in Education Research Journal, 27*(2), 188-200.

11. Vo, A., Le, M., & Nguyen, T. (2022). Mentorship and task relevance: Key factors in internship satisfaction and employability. *International Career Journal, 19*(5), 323-335.

12. ElSharkawy, A., Abdelrahman, N., & Soliman, H. (2023). Addressing biases in employability prediction: Advanced algorithmic approaches. *Journal of Higher Education and Data Analytics, 8*(3), 215-230.

13. Baker, D., Thompson, R., & Miller, K. (2023). Internships as a pathway to professional

success: An analytical perspective. *Workforce Development Studies, 17*(2), 165-180.

14. Casuat, J., Pérez, F., & Gonzalez, M. (2024). Tailoring employability models for diverse populations: A demographic perspective. *Regional Employment and Education Journal, 21*(1), 67-80.

15. Oberman, T., Weinstock, M., & Daniels, C. (2023). A holistic approach to internship program design: Insights from students and industry professionals. *Journal of Internship Studies, 13*(4), 298-312.