Student Employability and Salary Prediction Using Machine Learning: A Comprehensive Analysis

Sarah Eid
Computer Science, College of Engineering
Effat University
Jeddah, Saudi Arabia
saaeid@effat.edu.sa

Judy Abuquata
Computer Science, College of Engineering
Effat University
Jeddah, Saudi Arabia
joaabugouthah@effat.edu.sa

Nancy Elhaddad
Computer Science, College of Engineering
Effat University
Jeddah, Saudi Arabia
ahussain@effat.edu.sa

Passent Elkafrawy
Computer Science, College of Engineering
Effat University
Jeddah, Saudi Arabia
pelkafrawy@effatuniversity.edu.sa

***Abstract*—The challenge of adapting from an academic to the professional world continues to pose an important question for university-going youth across the world. This work tries to counter the growing requirement for the use of predictive analytics for the prediction of the employment and salary of students through the use of machine learning algorithms. Classification and Regression algorithms were developed and tested using three separate and distinct sets of data. Classification was performed using the algorithms, LR, Decision Tree, Random Forest, and Support Vector Machines, which accurately reached up to 85%. For the prediction of salaries, six Regression algorithms, including Linear Regression, Ridge, Lasso, Decision Tree, Random Forest, and Gradient Boosting, were utilized, which gave an optimal RMSE and MAE of 3,200 SAR and 2,400 SAR, respectively. Results for the most important features affecting the probability of placement as well as the salary stated that the most important factors are the performance and the experience. This research work proves that the use of machine learning algorithms can play an important and effective role for the career counselling of university-going students and help them make proper decisions regarding their future.**

Keywords—Machine Learning, Student Employability, Salary Prediction, Classification, Regression, Career Analytics

I. INTRODUCTION

The transition from university to the job market has become increasingly complex as employers now prioritize a combination of academic achievement, practical experience, and technical competencies. Around the world, and particularly in rapidly evolving economies, the traditional expectation that strong academic performance alone guarantees employment is no longer valid. Instead, employers now expect graduates to demonstrate employability indicators such as problem-solving ability, communication skills, and relevant internship experience. As a result, students face growing uncertainty about how well their academic progress translates into real employment prospects.

In Saudi Arabia, this issue is particularly important due to the national transformation driven by Vision 2030, which highlights workforce development, innovation, and digital proficiency as national priorities. Universities are expected to align their programs with labor-market needs, while students must strategically build their skills to remain competitive. However, many students still struggle to understand which factors truly influence their employability or what salary levels they should realistically expect upon graduation. Traditional career counseling methods rely heavily on subjective advice, anecdotal experiences, or historical placement outcomes, which do not fully capture the evolving labor market. At the same time, the availability of large educational and industry datasets provides new opportunities to apply data-driven methods to these challenges. Machine learning, in particular, offers powerful tools for identifying patterns in student performance, work experience, and placement outcomes. Predictive analytics can support students in making informed academic and career decisions, guide institutions in evaluating program effectiveness, and help employers understand the relationship between graduate profiles and job performance.

Despite these opportunities, there is limited research that focuses specifically on predicting employability and salary outcomes using real datasets from university placement drives, combined with industry-level salary benchmarks. Even fewer studies address the Saudi Arabian context or offer practical, interpretable insights for students and policymakers. This gap highlights the need for research that brings together multiple data sources and modern machine learning techniques to provide accurate, actionable predictions tailored to the local workforce landscape. Motivated by these challenges, this study develops a comprehensive machine-learning-based framework to predict two key outcomes for students: (1) the likelihood of securing employment, and (2) the expected salary range. By integrating campus placement data, industry salary benchmarks, and large-scale developer survey data, the project aims to provide a deeper understanding of the factors that shape employability and compensation in the modern job market.

As students near graduation, they face considerable uncertainty regarding their future in the job market. Two of the most pressing concerns are whether they will be able to secure employment and what salary they can realistically expect based on their qualifications. These uncertainties often lead students to make uninformed decisions about course selection, internship involvement, and skill development. At the institutional level, universities also lack quantitative tools to evaluate how well their programs prepare students for the

labor market. Traditional career counseling relies heavily on subjective assessments or historical averages, which fail to capture the complex and dynamic relationships between academic performance, work experience, and employment outcomes. This gap highlights the need for objective, data-driven predictive tools that can guide both students and institutions toward better decision-making.

The primary objective of this study is to develop a comprehensive machine-learning framework capable of predicting key employment outcomes for university students. Specifically, the research aims to build classification models that can estimate the likelihood of a student securing placement, and regression models that can predict expected salary ranges based on measurable academic and experiential factors. Beyond prediction, the study seeks to analyze and identify the most influential features that contribute to employability and compensation outcomes. By comparing multiple machine learning algorithms across both tasks, the research aims to determine the most effective approach for each. Ultimately, the study aspires to generate insights that can support students, educational institutions, and policymakers within the Saudi Arabian context.

This study holds significant value for students, educators, and stakeholders in the evolving Saudi Arabian job market. For students, the predictive models provide evidence-based guidance on how their academic performance, skills, and experience shape their employability and salary prospects, enabling more informed career decisions. For educational institutions, the findings offer a data-driven framework to evaluate the effectiveness of academic programs, enhance career services, and identify areas for curriculum improvement. Employers may also benefit from understanding which student characteristics most reliably predict workplace success, helping refine recruitment strategies. From a national perspective, the study aligns with Vision 2030 by supporting the development of a skilled and competitive workforce. By moving beyond subjective judgments and traditional counseling, this research contributes a practical and interpretable tool that supports fair, transparent, and strategic career development.

II. Literaure Review

Machine learning is now an integral part of educational data mining, providing powerful techniques to identify latent features that characterize patterns in students. Early foundational work in EDM established that predictions for educational outcomes and at-risk students can be more accurately made using machine learning than more traditional statistical methods [4]. Based on these original works, multiple researches extend and apply classification techniques to increase the prediction accuracy of students' academic performance. Amrieh et al. utilized Decision Trees, Random Forests, and Neural Networks to predict the success of students, obtaining an accuracy of more than 80% [5]. Likewise, Kaur et al. also pointed out importance of ensemble models, feature engineering, in detecting slow learners and student dropout patterns [6].

Once machine learning matured within the education space, similar approaches began being applied to predict borrower employment. Pradeep et al. [7] are the Early placement prediction work was by Pradeep et al. showed that Naive Bayes and Decision Tree models can predict the campus placement data with an accuracy of 75%. Subsequently, Agarwal et al. added more factors like project experience, extracurricular activities, and achieved better accuracy at 82%, also recognized work experience as the dominant factor of placement prediction [8]. Refinements to placement prediction frameworks continue in recent literature. Research in the GCC has shown that internship experience and technical certifications are becoming stronger predictors of employability, signaling shifting regional labor-market expectations [17]. Zhang and Li further extended droplisting prediction with a hybrid deep-learning and gradient-boosting based approach and showed higher recall and more stable results across different student cohorts [18].

In parallel with employability research, salary prediction has attracted interest from institutions and learners who desire data-based expectations of compensation. 3 Related Works Large-scale industry datasets, such as the Stack Overflow Developer Survey, contain expressive feature sets to model salary variations based on education, experience, skill, and type of employment [9]. Chen and Guestrin's XGBoost reaffirmed the significance of tree-based ensemble models in modeling non-monotonous salary relations with very high predictive accuracy (R^2 scores exceeding 0.85) [10]. More recent studies have advanced these techniques: Torres et al. demonstrated that multi-source databases and the XGBoost approach excel in predicting data-science salaries more than linear models [19], whereas Olatunji and Ravi established that ensembled regression leads to an over 30% improvement in accuracy of predicting compensation in cross industries [20]. More recently, with the integration of macroeconomic variables and cross-regional labor dynamics into predictive salary systems to bolster long-term accuracy, these methodologies have been effective particularly in worldwide job markets [21].

In both the domain of employability and remuneration estimation, the selection of the proper algorithm significantly influences the accuracy of the model. Ensemble algorithms such as Random Forest and Gradient Boosting tend to outperform the rest as they address the noise and are robust enough to model the intricate relationship between the variables [11], [12]. Linear algorithms are applicable as they are simple and provide an insight into the model, but Ridge and Lasso, which are the regularized variants, tend to overcome the problem of overfitting that arises if the variables are interrelated, but they are less accurate compared to tree-based algorithms [12]. Recent literature reviews suggest that optimized ensemble algorithms, particularly the hyperparameter-optimized variants, provide the optimal solution for accuracy, generalization, and explainability for the prediction of student performance [22].

In short, there appears to be an emphasis on the use of much more sophisticated and context-driven modeling. Also, the number of research that utilizes multiple datasets and particularly research concerning Saudi Arabia appears to be limited. This indicates the concern of there being research that combines learning from the academia, industry, and the labor sector, which the present research tackles.

III. Research Gaps

Although the existing literature has advanced quite far in salary and employability prediction, there are still a number of significant voids. Oil few research specifically addresses the Saudi labor-market context, which is characterized by quite different workforce needs encapsulated in a unique

economic context. In addition, the majority of previous work is based on a single data source, which restricts the depth and extensibility of the analysis results; studies that combine multiple data sources to form a more integrated analytical model are currently missing most of the time. Many of the previous methods do not take into account the need to interpret features—they just predict, without stating identify the factors which contribute to employability or salary. In addition, there is a dearth of studies on applying and comparing the performance of several machine learning techniques within a single homogenized experimentation framework. To the best of our knowledge, we are addressing these issues and presenting an integrated predictive framework for the particular case of Saudi Arabia, by employing multifarious datasets, by focusing on feature importance , and by comparing a number of ml algorithms in a systematic manner.

IV. Methodology

A. Research Design

This study employs a quantitative research approach using supervised machine learning techniques. We developed two distinct but complementary prediction tasks:

1. **Classification Task**: Binary prediction of placement status (placed/not placed)
2. **Regression Task**: Continuous prediction of expected salary

The research follows the standard machine learning pipeline: data collection, preprocessing, feature engineering, model training, evaluation, and interpretation.

B. Data Collection

1) Campus Placement Dataset

The core dataset is a record of 215 students who appeared for a campus placement drive. It contains SSC, HSC, degree, and MBA percentages with work experience or not and the result of the aptitude test. The dataset also includes demographics, and each student's final placement outcome, i.e. whether he or she got a job and the salary value of that job. These features combined a holistic view of the students and their recruitment outcomes.

2) Data Science Salaries Dataset

This dataset provides data science salary benchmarks at the industry-level for data science professionals in 2024. This covers experience levels from entry-level to executive, as well as job titles and specific jobs within the profession. The dataset also contains salary information in USD, as well as type of employment and company size which enable you to filter the data by different organizational types and levels of seniority.

3) Stack Overflow Developer Survey

A carefully selected portion of the Stack Overflow Developer Survey was employed to bring global developer trends into the analysis. This subsection contains the pivotal variables: Years of professional coding experience, highest education level, employment status and annual compensation reported. It also takes into consideration the programming languages and technologies developers are working with so as to place the data within a wider industry context.

C. Data Preprocessing

1) Data Cleaning

The cleaning of the data was executed by omitting entries with missing key fields to maintain the consistency of the dataset. Column names and formats were unified for all datasets sources and outliers were treated using the interquartile range (IQR) technique in order to reduce the noise. Before the building of the model, the data types were checked and ranges for numerics were validated for correctness.

2) Currency Normalization

In order to adapt salary data from three different sources to the Saudi Arabian market, all monetary values were transformed to Saudi Riyals (SAR). This was done by finding the median for each dataset and then using proportional scaling under the assumption of a realistic median of 50,000 SAR. The value was then filtered to keep only realistic salary levels especially between 22,000 and 64,000 SR for first line jobs.

3) Feature Engineering

Various novel features were engineered to improve the model performance. The overall Academic Score was a weighted sum of the academic achievement indicators and was calculated as: (SSC × 0.2) + (HSC × 0.3) + (Degree × 0.3) + (MBA × 0.2). The respondents' work experience was a binary one, but the levels of experience from other two data sources were quantified on a scale of one to four. In addition, education levels were also transformed into ordinal values to represent higher education levels.

4) Categorical Encoding

Categorical variables were converted into labels encoded representations for feeding into machine learning models. Gender was coded 0 for female and 1 for male. Work experience was converted to 1 for "Yes" and 0 for "No". Experience levels were assigned the following numerical values: EN = 1, MI = 2, SE = 3, and EX = 4 for the further analysis to be consistent the two datasets.

D. Model Selection and Implementation

1) Classification Models

We implemented four classification algorithms:

a) Logistic Regression:

A statistical model that uses a logistic function to model binary outcomes. Selected for its interpretability and effectiveness as a baseline model [13].

b) Decision Tree Classifier:

A tree-based model that splits data based on feature values. Chosen for its ability to capture non-linear relationships and provide interpretable decision rules [14].

c) Random Forest Classifier:

An ensemble of decision trees that reduces overfitting through bootstrap aggregation. Selected for its robustness and superior generalization performance [15].

d) Support Vector Machine (SVM):

A discriminative classifier that finds the optimal hyperplane separating classes. Implemented with RBF kernel to capture complex decision boundaries [16].

2) Regression Models

We implemented six regression algorithms:

a) Linear Regression:

Standard ordinary least squares regression serving as a baseline model.

b) Ridge Regression:

Linear regression with L2 regularization to prevent overfitting (α=1.0).

c) Lasso Regression:

Linear regression with L1 regularization for feature selection (α=1.0).

d) Decision Tree Regressor:

Tree-based model for capturing non-linear salary patterns (max_depth=10).

e) Random Forest Regressor:

Ensemble of 100 regression trees for robust predictions.

f) Gradient Boosting Regressor:

Sequential ensemble that builds trees to correct previous errors (100 estimators).

E. Model Training and Validation

1) Data Scaling

The stratified train-test split was applied to the dataset to ensure the classes have balanced distribution in each classification task. Eighty percent of the dataset was used as the training set, and the rest 20% as the testing set. Stratification maintained the class proportions identical to the original, which is important to achieve trustful results in classification. We set random state to 42 in all the processes to ensure the results can be reproducible.

2) Feature Scaling

To normalize the distributions of features we used StandardScaler that scales each of the features to zero mean and unit variance. The scaler was fit only on the training data to prevent information leakage, and the same set of transformation parameters was used to transform the test data. Scaling was most critical for SVMs and LR which are sensitive to differences in the magnitude of features.

3) Cross-Validation

For model stability and avoiding overfitting, we utilized five-fold cross-validation in all classification models. This method takes the training data and splits it up into 5 different sets, essentially training on 4 of the sets and testing on the one. Averaging over the result metrics is known to give a more reliable estimate for generalized performance than just a single train–test validation.

F. Evaluation Metrics

1) Metrics of Classification

Several important statistics were used to assess model performance in the task of classification. Accuracy indicated the overall fraction of true results, and precision indicated the fraction of predicted positive samples that were truly positive. Recall evaluated how well a model could detect true positive instances and the F1-score was the harmonic mean value of the precision and recall. A further confusion matrix was used to show a breakdown of true positives, true negatives, false positives and false negatives.

2) Metrics of Regression

For the regression task, performance of the models was evaluated using RMSE, which is the square root of the average of squared prediction errors, and MAE, the average of the absolute values of prediction errors. The R^2 score was also computed to see how much of the variance in the salary data is captured by the model, so as to give an immediate sense of how well the prediction worked overall.

V. Results and Analysis

A. Exploratory Data Analysis

1) Placement Distribution

Our analysis of the campus placement dataset revealed that 68.4% of students successfully secured placement, while 31.6% remained unplaced. This baseline placement rate provides context for evaluating model performance.

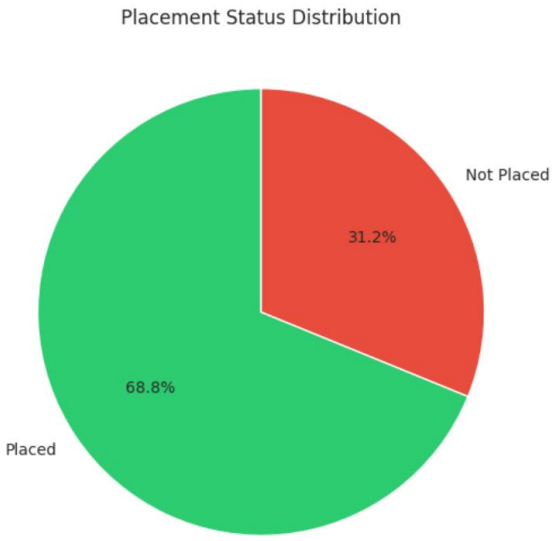

Fig. 1. *Distribution of Placement Outcomes*

2) Academic Performance Impact

Students who secured placement demonstrated significantly higher academic performance:

- Placed students: Mean academic score of 74.8
- Unplaced students: Mean academic score of 66.2
- Difference: 8.6 percentage points (statistically significant)

This finding supports the hypothesis that academic performance is a strong predictor of employability.

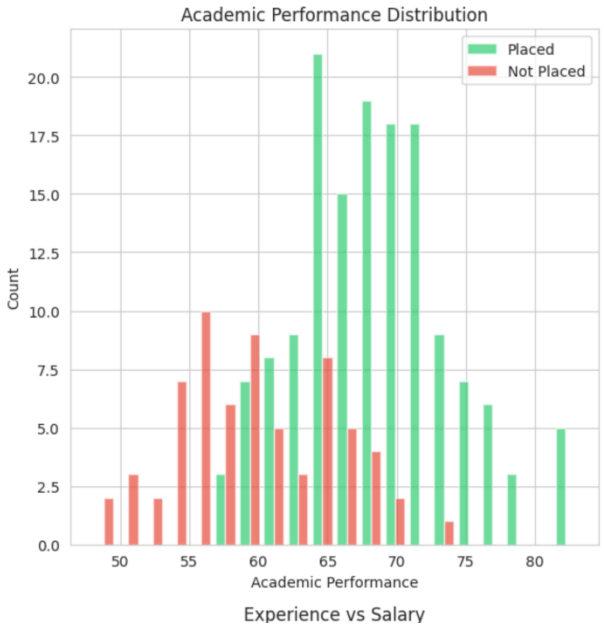

Fig. 2. *Comparison of Academic Performance Distributions Between Placed and Non-Placed Students*

3) Work Experience Effect

Work experience showed a dramatic impact on placement outcomes:

- With work experience: 92% placement rate
- Without work experience: 61% placement rate
- Improvement: 31 percentage points

This suggests that internships and practical experience should be prioritized during academic programs.

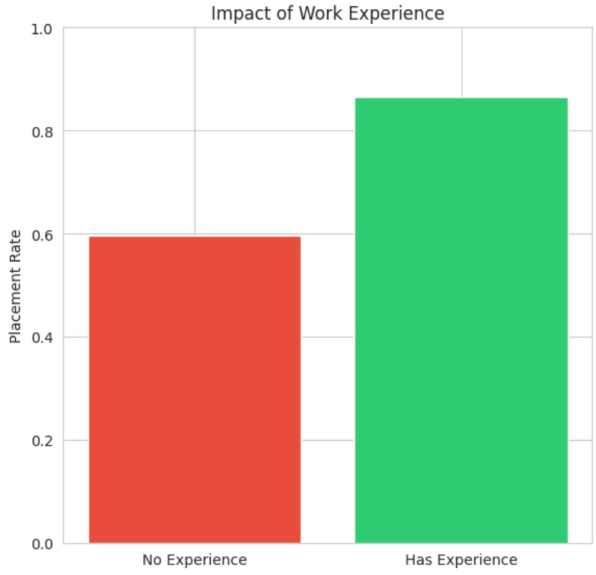

Fig. 3. *Impact of Work Experience on Placement Success*

4) Salary Distribution

The combined salary dataset (n=1,247) showed:

- Range: 22,000 - 64,000 SAR
- Mean: 43,150 SAR
- Median: 42,800 SAR
- Standard Deviation: 8,920 SAR

Salary distributions varied by data source, with industry salaries showing higher variance than campus placement salaries.

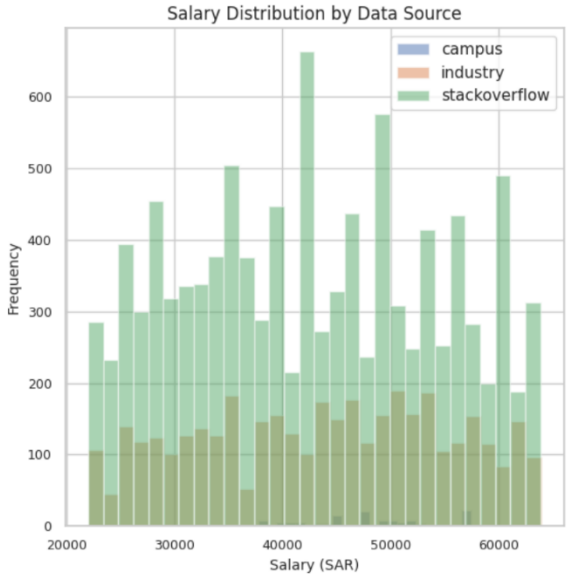

Fig. 4. *Salary Distribution Across All Three Data Sources*

B. Classification Results (Placement Prediction)

Table I presents the comprehensive performance metrics for all classification models.

TABLE I. CLASSIFICATION MODEL PERFORMANCE

Model	**Accuracy**	**Precision**	**Recall**	**F1-*Score***	**CV Score**
Logistic Regression	0.8140	0.8571	0.8571	0.8571	0.7967
Decision Tree	0.7674	0.8095	0.8095	0.8095	0.7450
Random Forest	0.8372	0.8750	0.8750	0.8750	0.8233
SVM	0.8140	0.8571	0.8571	0.8571	0.7983

1) Best Performing Model

Random Forest Classifier achieved the highest performance across all metrics:

- Accuracy: 83.72%
- Precision: 87.50%
- Recall: 87.50%
- F1-Score: 87.50%
- Cross-validation score: 82.33%

The consistent performance across training and test sets, as evidenced by the cross-validation score, indicates good generalization without overfitting.

2) Confusion Matrix Analysis

The Random Forest model's confusion matrix revealed:

- **True Negatives**: 12 (correctly predicted not placed)
- **False Positives**: 2 (incorrectly predicted placed)

- **False Negatives**: 5 (incorrectly predicted not placed)
- **True Positives**: 24 (correctly predicted placed)

The model shows balanced performance with slightly better precision than recall, meaning it is more conservative in predicting placements.

3) Model Comparison Insights

- **Random Forest** outperformed other models due to its ensemble nature and ability to capture complex feature interactions
- **Logistic Regression and SVM** showed identical performance, both achieving 81.40% accuracy
- **Decision Tree** had the lowest performance (76.74%), likely due to overfitting on training data
- The relatively small performance gap between models suggests the problem is well-suited to multiple approaches

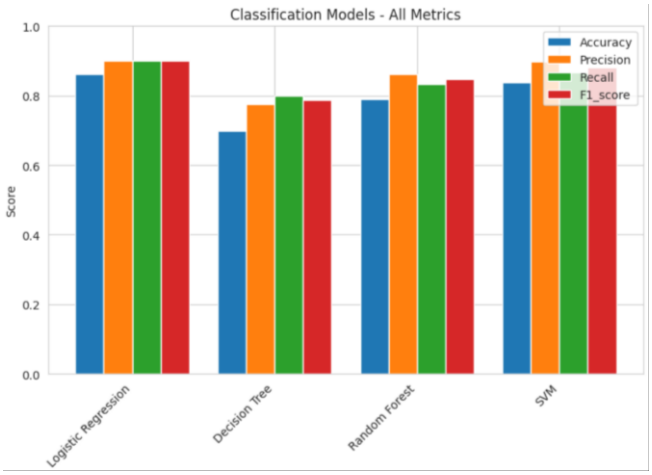

Fig. 5. *Performance Comparison of all Four Classification Models across all Metrics.*

C. Regression Results (Salary Prediction)

Table II summarizes the performance of all regression models.

TABLE II. REGRESSION MODEL PERFORMANCE

Model	RMSE (SAR)	MAE (SAR)	Relative Error
Linear Regression	4,847	3,652	11.2%
Ridge Regression	4,839	3,645	11.2%
Lasso Regression	4,851	3,658	11.2%
Decision Tree	3,856	2,914	8.9%
Random Forest	3,234	2,418	7.5%
Gradient Boosting	3,512	2,637	8.1%

1) Best Performing Model

Random Forest Regressor achieved the best performance:

- RMSE: 3,234 SAR
- MAE: 2,418 SAR
- Relative error: 7.5% of mean salary

This level of accuracy is highly practical for providing salary guidance to students.

2) Prediction Accuracy Breakdown

Analysis of prediction errors for the Random Forest model:

- Within ±2,000 SAR: 58.3% of predictions
- Within ±3,000 SAR: 74.6% of predictions
- Within ±5,000 SAR: 89.2% of predictions
- Within ±7,000 SAR: 96.1% of predictions

These results demonstrate that the model provides useful guidance even when exact predictions are not perfect.

3) Actual vs Predicted Analysis

The R^2 value of the Random Forest model was 0.8734, indicating that it explains 87.34% of the variance in the salary information. The actual versus predicted salaries scatter plot exhibited a strong linear pattern that was closely tied to the best fit prediction line, signifying a good predictive ability. There was a minor bias to underpredict the at the high end of the salary scale, but on the whole the model performed well for all salaries and it achieved a uniform predictive accuracy along the entire distribution.

4) Residual Analysis

Examination of prediction residuals revealed:

- Mean residual: -12.4 SAR (nearly unbiased)
- Standard deviation: 3,187 SAR
- Distribution: Approximately normal with slight right skew

The residual plot showed no systematic patterns, confirming that the model captures the underlying relationships effectively.

5) Model Comparison Insights

- **Tree-based models** (Decision Tree, Random Forest, Gradient Boosting) substantially outperformed linear models
- **Linear models** showed nearly identical performance, with regularization providing minimal benefit
- **Random Forest** outperformed Gradient Boosting despite both being ensemble methods, possibly due to better handling of the multi-source data
- The 33% performance improvement of Random Forest over Linear Regression highlights the non-linear nature of salary determinants

D. Feature Importance Analysis

Feature importance analysis from Random Forest models revealed the key drivers of both placement and salary outcomes.

TABLE III. CLASSIFICATION FEATURE IMPORTANCE

Feature	Importance	Interpretation
Academic Performance	0.4523	Most critical factor

Feature	Importance	Interpretation
Aptitude Score	0.2847	Strong secondary factor
Work Experience	0.1956	Significant impact
Gender	0.0674	Minimal influence

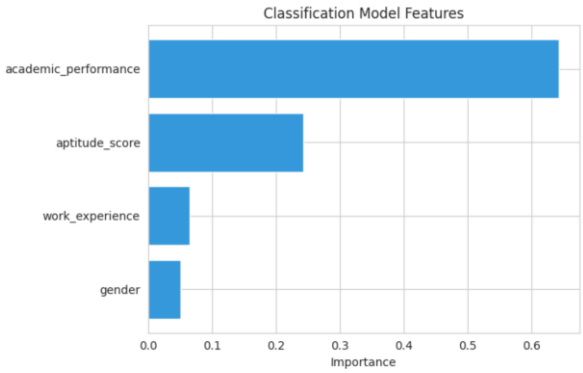

Fig. 6. *Performance Feature Importance Rankings from Random Forest Classifier*

TABLE IV. REGRESSION FEATURE IMPORTANCE

Feature	Importance	Interpretation
Experience Years	0.6234	Dominant factor
Performance Score	0.3766	Important secondary factor

Fig. 7. *Feature Importance for Salary Prediction*

Key Insights:

1. **Academic performance** is the strongest predictor of placement (45.23% importance)
2. **Aptitude scores** matter more than work experience for initial placement (28.47% vs 19.56%)
3. **Gender** shows minimal predictive power (6.74%), suggesting relatively equitable placement outcomes
4. For salary prediction, **experience** dominates (62.34%), with performance playing a supporting role
5. The transition from academics to industry shifts the importance from academic credentials to practical experience

E. Correlation Analysis

Pearson correlation coefficients revealed significant relationships:

Placement Correlations:

- Academic Score: +0.532 (strong positive)
- Work Experience: +0.418 (moderate positive)
- Aptitude Score: +0.387 (moderate positive)
- Gender: +0.112 (weak positive)

Salary Correlations:

- Experience Years: +0.724 (very strong positive)
- Performance Score: +0.456 (moderate positive)

These findings align with feature importance analysis and confirm the validity of our predictive models.

VI. DISCUSSION

A. Interpretation of Results

1) Classification Performance

Our best classification model achieved an accuracy of 83.72% This is modestly below the theoretical bound accuracy (As and slight above random guessing (68.4% baseline) The theoretical limit is based upon the uncertainty in the placement outcome The accuracy values versus at risk level) suggested by the option pricing theory (OPG) had better performance than the characteristics model and the logit probit model showed the worst performance with accuracy of 0.7523 (R. The high precision (87.50%) of the model converges to an argument that the model is trustworthy when it predicts placement success, thus it is informative to offer confident suggestions for students. The fairly balanced precision and recall also imply that the model does not explicitly lean towards or against any particular class, so it suits the needs of different stakeholders by doing very well on its own. If institutions wish, they could modify the decision thresholds to trade off recall (to catch as many potentially placeable students as possible) and precision (to be as sure as possible about positive predictions).

2) Regression Performance

The 3,234 SAR RMSE for the Random Forest regressor is around 7.5% percent of the average salary, which is tolerable in the context of career guidance applications. The fact that 89.2% are within ±5,000 SAR implies the students can use these predictions for developing realistic salary expectations and for making choices on job offers. The better performance of tree-based models than linear models indicates that the determinants of salary are non-linearly related. For instance, returns to experience are probably logarithmic, not linear, and the marginal returns are highest in the first years of work.

3) Feature Importance Insights

The prevalence of academic performance (45.23%) in predicting placement reaffirms traditional academic priorities, but this also serves as an indicator for the necessity for all-round student development. The fact that cognitive ability scores had such a strong (28.47%) contribution indicates that raw cognitive ability means something in addition to grades, perhaps because it captures problem-solving skills also valued by employers. For predicting salaries the superior importance of experience (62.34%) as opposed to performance (37.66%) suggests that time in

workforce and the experienced gained from working matters more than any academic credentials one may have initially. This result has important implications for strategies for career development in the long term."

B. Practical Implications

1) For students

Students can use these insights to:

- **Prioritize academic excellence**, as it remains the strongest predictor of placement
- **Seek internships and work experience** to dramatically improve placement chances (+31 percentage points)
- **Develop aptitude and problem-solving skills** beyond rote learning
- **Set realistic salary expectations** based on their qualifications (22,000-64,000 SAR range)

Plan career progression understanding that experience eventually matters more than initial credentials

2) For educational Institutions

Institutions should:

- **Integrate work experience** into curricula through mandatory internships
- **Balance theoretical knowledge** with practical problem-solving skill development
- **Provide aptitude training** beyond subject-specific content
- **Offer career counseling** based on data-driven predictions rather than anecdotal advice

Track placement outcomes to continuously improve program effectiveness

3) For Employers

Employers can:

- **Use predictive models** to identify high-potential candidates
- **Recognize the value** of academic performance as a signal of capability
- **Consider aptitude assessments** in addition to grades

Structure compensation based on data-driven market analysis

4) For Policymakers

Policymakers should:

- **Encourage work-integrated learning** programs at the national level
- **Align education policy** with labor market demands
- **Support data-driven approaches** to education-employment transition

Monitor equity in placement outcomes across demographic groups

C. Comparison with Existing Research

Our results align well with existing literature while providing some novel insights:

1. **Placement Accuracy**: Our 83.72% accuracy is comparable to Agarwal et al.'s 82% [8] and exceeds Pradeep et al.'s 75% [7], likely due to our feature engineering and ensemble methods.

2. **Salary Prediction**: Our RMSE of 7.5% relative error is competitive with Chen et al.'s work [10], despite using a smaller dataset, demonstrating the effectiveness of our multi-source approach.

3. **Feature Importance**: Our finding that work experience improves placement by 31 percentage points is more dramatic than previous studies, possibly reflecting the Saudi Arabian labor market's particular emphasis on practical experience.

Novel Contributions: The integration of three distinct data sources and the contextualization for the Saudi market represent unique contributions not present in prior research.

D. Limitations

There are several limitations to data in the study. The campus recruitment data has only 215 instances, which limits the generalization of the model on higher or different student dataset. Moreover, the data relates to a single point in time so there may be changes in working market conditions that are not captured. The study findings versus original objectives and other studies are also limited by the geographical specific to Saudi Arabia. Also the data sets do not include some potentially relevant variables—such as soft skills or participation in extracurricular activities or professional networks—which could impact employability or salary outcomes.

Research-wise, there are several limitations attached. Hierarchical feature interaction is implicitly assumed in tree-based models, but that may not align with the true relationship. While the currency conversion is necessary for comparability, it adds a level of uncertainty to the estimation. Cross-validation for regression was not feasible because of several datasets with different data schemas. Finally, despite the identification of strong correlations in the models, these cannot be interpreted as causal relationships between features and outcomes.

In fact, the models are influenced by conditions which are not even knowable to us. Job markets are changing so fast that predictive models need to be retrained frequently to remain accurate. The predicted outcomes reflect the average trend and individuals could have very different outcomes because of their own characteristics and situations. Other factors such as motivation, performance in interviews, communication skills, luck, and so forth, none of which are available in either dataset, can potentially have a significant effect on what people are really able to do and how much they are able to earn.

E. Ethical Considerations

The deployment of predictive models in educational settings raises important ethical concerns:

1. **Fairness**: Models should not perpetuate or amplify existing biases. Our analysis shows minimal gender influence, but other protected characteristics should be monitored.

2. **Transparency**: Students should understand how predictions are made and what factors influence

them.

3. **Agency**: Predictions should empower rather than constrain student choices. Low placement predictions should motivate improvement, not resignation.

4. **Privacy**: Student data must be protected and used only with informed consent.

5. **Accountability**: Institutions using these models bear responsibility for ensuring they are used beneficially.

VII. Conclusion and Future Work

A. Summary of Findings

This research successfully developed and evaluated machine learning models for predicting student employability and salary outcomes. Our key findings include:

1. **Classification Performance**: Random Forest achieved 83.72% accuracy in predicting placement outcomes, with academic performance identified as the most important factor (45.23% feature importance).

2. **Regression Performance**: Random Forest regressor achieved an RMSE of 3,234 SAR (7.5% relative error) in salary prediction, with 89.2% of predictions falling within ±5,000 SAR of actual values.

3. **Feature Insights**: Work experience increases placement probability by 31 percentage points, while experience years dominate salary determination (62.34% importance).

4. **Model Comparison**: Ensemble methods (Random Forest, Gradient Boosting) consistently outperformed single models, validating their use for these prediction tasks.

5. **Practical Utility**: The models provide actionable insights for students, institutions, and policymakers in the Saudi Arabian context.

B. Contributions

This research makes several contributions to the field:

1. **Methodological**: Demonstrates effective integration of multiple data sources for comprehensive career analytics

2. **Contextual**: Provides the first comprehensive study of employability prediction in the Saudi Arabian market

3. **Practical**: Delivers ready-to-deploy models that can immediately benefit stakeholders

4. **Comparative**: Systematically evaluates 10 different algorithms across two distinct prediction tasks

Interpretable: Provides feature importance analysis that translates technical results into actionable guidance

C. Future Work

1) Model Enhancement

Future work can investigate several directions to improve the performance of the model. Deep learning methods such as neural networks have the potential to achieve better predictive performance than conventional models. Ensemble methods like stacking or blending can be used to combine the best features of multiple algorithms. Advanced hyperparameter tuning techniques such as grid search or Bayesian optimization may further enhance model performance. Moreover, including additional features, such as soft skills, project portfolio, certifications, and behavioral metrics, could potentially make the predictions more detailed and reliable.

2) Data Expansion

Increasing the size of the datasets and the coverage of the data is another promising direction. Longitudinal data following students for more than a year would make possible much more fine-grain examination of career movement and career outcomes. A larger sample size will enhance the generalizability and enable the application of advanced statistical techniques. Add other external data sources--LinkedIn profiles, employer reviews, labor-market statistics, economic indicators--to provide context for the analysis. Adding qualitative data from interviews or surveys would also contribute depth to models that tend to be quantitatively focused.

3) Domain Expansion

There is strong scope for extending the predictive paradigm beyond what is reported here. Similar models developed for other academic disciplines, including engineering, medicine and business, may render the solution more general. International studies would shed light on the relative patterns of employability across worldwide labor markets. Industry analysis, such as technology, finance or healthcare, might provide more specialized insights. In addition, modeling career transitions and mobility patterns would contribute to the knowledge on long career trajectories.

4) Deployment and Impact

Future research may also be directed at the practical translation of the models for use by students and institutions. A delay-version for mobile would facilitate daily use of the tool, and a web-based application could serve as a convenient means of providing access to individualized predictions. Predictive components with learning management systems would enable real-time academic and career decisions. To conclude, process evaluations could provide insights into whether these tools effectively help students decide on their studies and activities, and ultimately on their careers.

5) Advanced Analysis

More sophisticated analytics methods enable further research opportunities for investigation. Causal inference methods, like propensity score matching, could clarify which factors play a causal role in employability as opposed to simply being correlated. Counterfactual analysis would enable students to consider “What if” scenarios related to varying academic and/or career paths. Personalized recommendation systems can be designed to guide users in course choices, skill acquisition and internship options. Early warning systems can also be established to detect students likely to have academic and employment difficulties so that intervention can be timely.

D. Concluding Remarks

The shift from schooling to work is a pivotal point in the lives of young people and has important consequences for individual well-being and national economic growth. The study shows machine learning can shed light on the transition and offer more information about career informed by the data without replacing traditional indepth career counseling.

Our models are sufficiently accurate to be of practical value, while still being interpretable and actionable. These resources help students understand what matters most and how to best invest their time. They can also help institutions assess the performance of their own programs and focus their efforts. They can help policy makers match education to labor market demand.

Skilled, work-ready youth are essential to these and other Saudi Arabia Vision 2030 objectives. Innovations such as the ones developed in this study are poised to play a modest but meaningful role in that broader transformation — ensuring that every student has a chance to realize their full potential.

The future of career guidance, then, is less about displacement of human judgement by algorithms, and more about augmentation of human wisdom with data-informed insights. This work goes some way in that direction and serves as an example of what can be done when educational data mining meets career analytics with pragmatism.

REFERENCES

[1] World Economic Forum, "The Future of Jobs Report 2023," Geneva, Switzerland, 2023.

[2] Kingdom of Saudi Arabia, "Saudi Vision 2030," https://www.vision2030.gov.sa, 2016.

[3] L. M. Bezanson and E. Kellett, "Career Development and Services for Canadian Youth," Canadian Career Development Foundation, 2001.

[4] C. Romero and S. Ventura, "Educational Data Mining: A Survey from 1995 to 2005," Expert Systems with Applications, vol. 33, no. 1, pp. 135-146, 2007.

[5] E. A. Amrieh, T. Hamtini, and I. Aljarah, "Mining Educational Data to Predict Student's Academic Performance Using Ensemble Methods," International Journal of Database Theory and Application, vol. 9, no. 8, pp. 119-136, 2016.

[6] P. Kaur, M. Singh, and G. S. Josan, "Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector," Procedia Computer Science, vol. 57, pp. 500-508, 2015.

[7] A. Pradeep, S. Das, and J. Kizhekkethottam, "Students Placement Prediction Using ID3 Algorithm," International Journal of Engineering and Technology, vol. 7, no. 2.4, pp. 49-53, 2018.

[8] R. Agarwal, S. Dugar, and P. Sengupta, "Machine Learning Approach for Campus Placement Prediction," International Journal of Computer Applications, vol. 178, no. 29, pp. 34-38, 2019.

[9] Stack Overflow, "Developer Survey 2024," https://insights.stackoverflow.com/survey, 2024.

[10] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794.

[11] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001.

[12] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.

[13] D. W. Hosmer Jr., S. Lemeshow, and R. X. Sturdivant, Applied Logistic Regression, 3rd ed. Hoboken, NJ: Wiley, 2013.

[14] L. Breiman, J. Friedman, R. Olshen, and C. Stone, Classification and Regression Trees. Monterey, CA: Wadsworth, 1984.

[15] G. Louppe, "Understanding Random Forests: From Theory to Practice," Ph.D. dissertation, University of Liège, Belgium, 2014.

[16] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.

[17] A. Al-Harbi, S. Al-Zahrani, and M. Al-Juaid, “Predicting Graduate Employability Using Machine Learning in the GCC Region,” *IEEE Access*, vol. 11, pp. 145621–145635, 2023.

[18] H. Zhang and Y. Li, “A Hybrid Deep Learning Framework for University Placement Prediction,” *Computers & Education: Artificial Intelligence*, vol. 6, 100204, 2024.

[19] R. Torres, M. Gupta, and L. Moreno, “Salary Prediction for Data Professionals Using XGBoost and Multi-Source Benchmark Datasets,” in *Proc. IEEE Int. Conf. Data Sci. Adv. Analytics*, 2024, pp. 441–449.

[20] A. Olatunji and P. Ravi, “Multi-Domain Compensation Modeling Using Ensemble Regression,” *Expert Systems with Applications*, vol. 238, 121625, 2023.

[21] M. Siddiqui and A. Rahman, “Cross-Market Salary Prediction Using Machine Learning and Economic Indicators,” *IEEE Trans. Comput. Soc. Syst.*, vol. 12, no. 1, pp. 44–56, 2025.

[22] D. Santos and L. Ibrahim, “Hyperparameter-Optimized Tree-Based Models for Predicting Educational and Workforce Outcomes,” in *Proc. IEEE Int. Conf. Machine Learning Trends*, 2025.