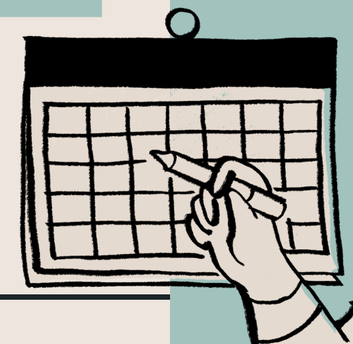
Data Science - Final Project

STUDENT EMPLOYABILITY AND SALARY PREDICTION USING MACHINE LEARNING

Sarah Eid | Judy Abuquata | Nancy Elhaddad

TABLE OF CONTENTS



PROBELM STATEMENT

THE CHALLENGE STUDENTS FACE

Two Critical Uncertainties:

- Will I get placed after graduation?
- What salary should I expect?

Current Issues:

- Traditional career counseling relies on subjective advice
- Students make uninformed decisions about courses and internships
- Institutions lack quantitative tools to evaluate program effectiveness

Our Solution:

DATA-DRIVEN PREDICTIVE MODELS USING MACHINE LEARNING

RESEARCH OBJECTIVES

Create Regression Models

Estimate expected salary ranges based on qualifications

Provide Actionable Insights

Generate practical guidance for Saudi Arabian context

Build Classification Models

Predict likelihood of student placement (Placed/Not Placed)

Identify Key Factors

Determine which features most influence employability and salary

Compare Algorithms

Evaluate different ML models to find the best approach

DA
TA
SE
TS

DATASETS OVERVIEW

DATASET	SIZE	KEY FEATURES
Campus Placement	215 students	Academic scores, aptitude, work experience, placement outcomes
Data Science Salaries	Industry benchmarks	Experience levels, job titles, salary data (USD)
Stack Overflow Survey	Developer insights	Years of experience, education, compensation, technologies

Why Multiple Sources?

- Comprehensive analysis combining academic, industry, and global perspectives
- More robust predictions through data integration
- Realistic salary benchmarks for Saudi market (normalized to SAR)

Campus Recruitment. (n.d.). Retrieved from www.kaggle.com website: https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement compatibility
Data Science Job Salaries. (n.d.). Retrieved from www.kaggle.com website: https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries
Stack Overflow Insights - Developer Hiring, Marketing, and User Research. (n.d.). Retrieved from survey.stackoverflow.co website: https://survey.stackoverflow.co/

METHODOLOGY OVERVIEW

1 - Data Collection → **2 - Preprocessing** → **3 - Feature Engineering**

↓

4 - Train/Test Split ←  ← **5 - Model Training**

↓

6 - Model Evaluation → **7 - ML Algorithms**

Key Steps:

- Data cleaning & outlier removal
- Currency normalization to SAR (22,000 - 64,000 range)
- Feature engineering (Academic Score, Experience encoding)
- 80/20 train-test split with cross-validation
- Standardization for model compatibility

CAMPUS PLACEMENT DATASET

Main Probelms:

- Salary provided in INR (not relevant for KSA context)
- Placement status was a text label
- Academic performance split into many separate percentage columns
- Gender and work experience were categorical strings
- Missing values in academic fields
- Some duplicated rows

Fixes:

- Converted salary from INR → SAR
- Created a binary placement label (placed_binary)
- Engineered a weighted academic_score from SSC/HSC/Degree/MBA
- Label-encoded gender and converted work experience
- Removed duplicates & filled missing academic fields using median

PREPROCESSING

STACK OVERFLOW DEVELOPER SURVEY

Main Probelms:

- Too many irrelevant columns in original dataset
- Missing compensation values (key to analysis)
- Years of coding contained non-numeric text values
- Programming languages stored as long text lists
- Salary values had a huge numeric range

Fixes:

- Kept only important columns (education, experience, salary, tech skills)
- Dropped rows missing compensation
- Converted experience field to numeric
- Counted number of languages from text
- Normalized compensation values

PREPROCESSING

GLOBAL SALARIES DATASE

Main Probelms:

- Duplicate records and missing salary entries
- Job titles inconsistent in formatting
- Experience level was categorical text (Entry/Mid/Senior/etc.)
- Salary in mixed currencies
- Remote/on-site environment hidden in text
- Company size categories too verbose

Fixes:

- Removed duplicates + dropped rows missing salary
- Normalized job titles to lowercase/trimmed
- Mapped experience levels to numeric values
- Converted salary to SAR based on dataset median
- Created remote_flag as 0/1
- Simplified company size into S/M/L categories

PREPROCESSING

Creating Meaningful Predictors

Academic Score (Weighted Average):

(SSC × 0.2) + (HSC × 0.3) + (Degree × 0.3) + (MBA × 0.2)

Categorical Encodings:

- Work Experience: Yes = 1, No = 0
- Gender: Male = 1, Female = 0
- Experience Level: EN=1, MI=2, SE=3, EX=4

Why This Matters:

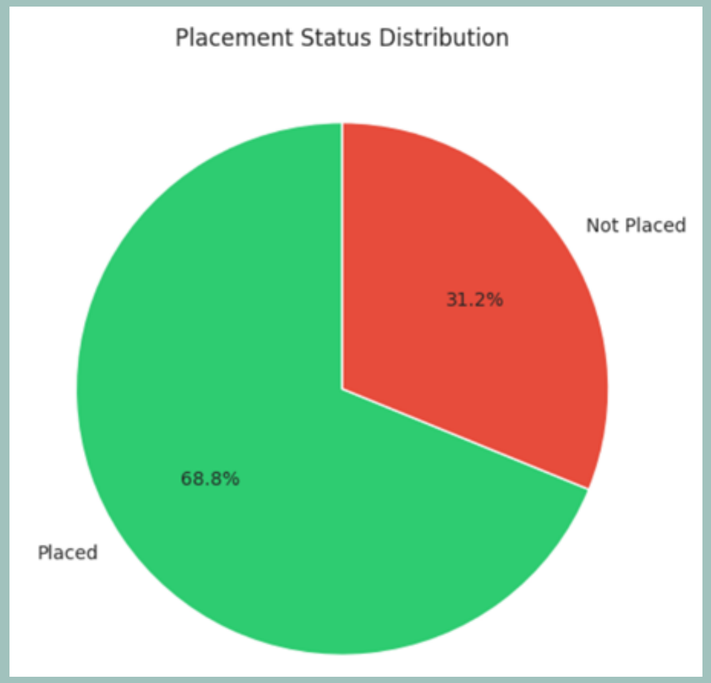
- Combines multiple academic metrics into single performance indicator
- Enables algorithms to process categorical variables
- Standardizes features across different data sources

FEATURE ENGINEERING

E D A

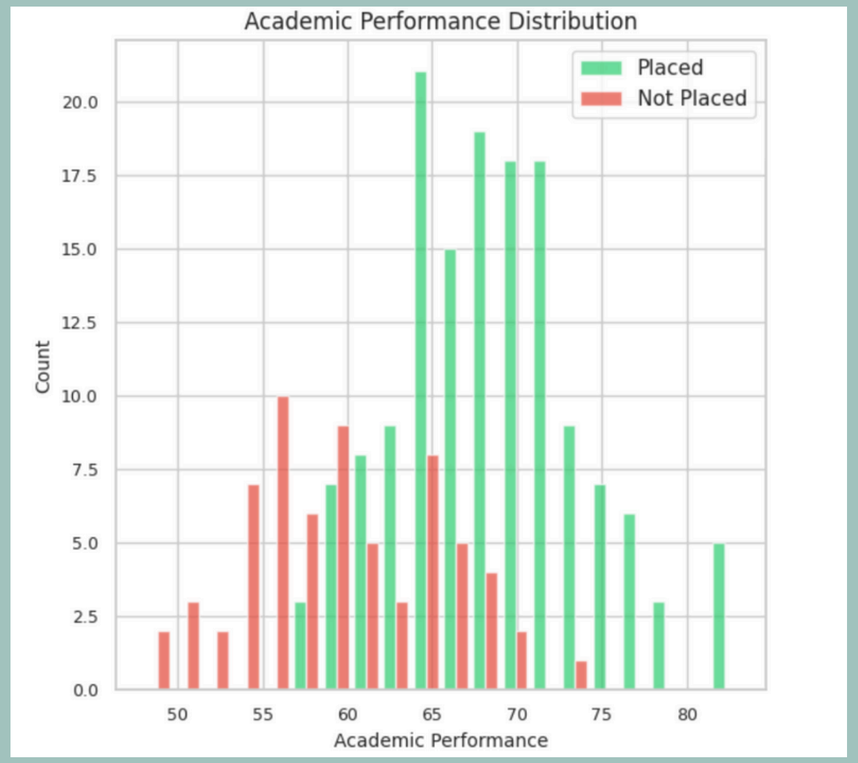
EXPLORATORY DATA ANALYSIS - KEY INSIGHTS

Placement Rate:



68.4% placed, 31.6% not placed

Academic Performance Impact:



- Placed students: Mean score 74.8
- Not placed students: Mean score 66.2
- Difference: 8.6 points

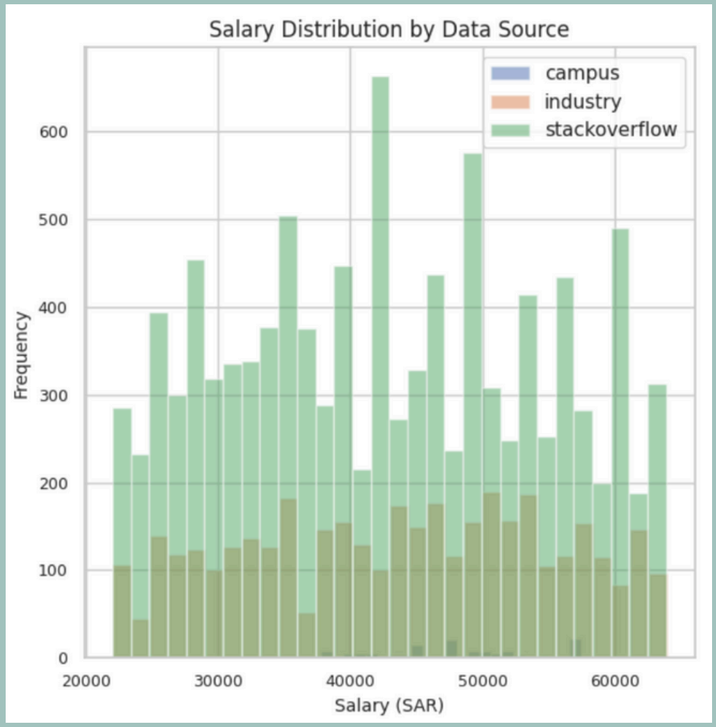
E D A

EXPLORATORY DATA ANALYSIS - KEY INSIGHTS

Work Experience Effect:

- With experience: 92% placement rate
- Without experience: 61% placement rate
- Impact: +31 percentage points

Salary Distribution:



- Range: 22,000 - 64,000 SAR
- Mean: 43,150 SAR | Median: 42,800 SAR

MACHINE LEARNING MODELS

Classification (Placement Prediction):

- Logistic Regression - Simple, interpretable baseline
- Decision Tree - Captures non-linear patterns
- Random Forest - Ensemble for robustness
- SVM (RBF kernel) - Complex decision boundaries

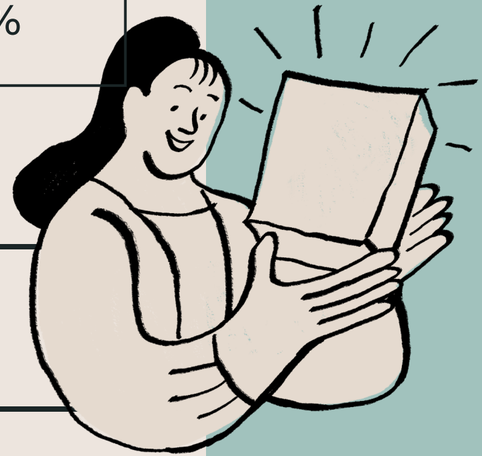
Regression (Salary Prediction):

- Linear Regression - Baseline model
- Ridge Regression - L2 regularization
- Lasso Regression - L1 regularization
- Decision Tree - Non-linear patterns
- Random Forest - Ensemble approach
- Gradient Boosting - Sequential error correction

CLASSIFICATION RESULTS

Placement Prediction Performance

MODEL	ACCURACY	PRECISION	RECALL	F1-SCORE
Logistic Regression	81.40%	85.71%	85.71%	85.71%
Decision Tree	76.74%	80.95%	80.95%	80.95%
Random Forest	**83.72%**	**87.50%**	**87.50%**	**87.50%**
SVM	81.40%	85.71%	85.71%	85.71%



REGRESSION

Salary Prediction 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%

RESULTS

For Placement Prediction:

Academic Performance: 45.23% - Most critical factor

Aptitude Score: 28.47% - Strong secondary predictor

Work Experience: 19.56% - Significant impact

Gender: Minimal influence (equitable outcomes

For Salary Prediction:

Experience Years: 62.34% - Dominant factor

Performance Score: 37.66% - Supporting role

Key Insight: Academic credentials matter most for placement, but experience dominates salary determination in the workforce.

FEATURE IMPORTANCE ANALYSIS

Work experience increases placement rate by 31 percentage points

Internships should be mandatory in curricula

Academic performance is the strongest placement predictor (45.23%

Excellence in academics remains crucial

Experience drives salary more than credentials (62.34% importance)

Long-term career growth depends on practical experience

Models are practically useful

89.2% of salary predictions within ±5,000 SAR

Ensemble methods outperform single algorithms

Random Forest best for both tasks (83.72% & 7.5% error)

KEY FINDINGS SUMMARY

PRACTICAL

PRACTICAL IMPLICATIONS

For Students:

- Prioritize academic excellence (strongest placement predictor)
- Seek internships/work experience (+31% placement boost)
- Develop problem-solving skills beyond rote learning
- Set realistic salary expectations (22K-64K SAR range)

For Institutions:

- Integrate mandatory internships into curricula
- Balance theory with practical skill development
- Implement data-driven career counseling
- Track placement outcomes for continuous improvement

For Employers:

- Use predictive models to identify high-potential candidates
- Value academic performance as capability signal
- Structure compensation based on market data

CONCLUSION

Technical Success:

- Random Forest: 83.72% placement prediction accuracy
- Random Forest: 7.5% relative error in salary prediction
- Systematic evaluation of 10 ML algorithms

Practical Value:

- Actionable insights for students, institutions, policymakers
- First comprehensive employability study for Saudi market
- Ready-to-deploy predictive framework

Key Takeaway: Machine learning can effectively support career counseling and help students make informed decisions about their academic and professional futures.

INTERFACE

Powered by Best-Performing ML Models

📊 Overview | 🎯 Live Predictor | 📈 Model Performance | ⬜ Batch Prediction

🎯 Live Prediction Tool

🏆 Using Best Models: Random Forest (88% accuracy) + Gradient Boosting (7K RMSE)

📝 Student Information

🎯 Prediction Results

📚 Academic Performance

10th Grade Percentage

75.00 − +

12th Grade Percentage

72.00 − +

Degree Percentage

70.00 − +

🎓 MBA Information

Have you completed MBA?

DEMO

THANK YOU
FOR LISTENING!