LAGUNA STATE POLYTECHNIC UNIVERSITY

College of Engineering

AI BOARD EXAM PREDICTION SYSTEM

Complete Machine Learning Training Report

Report Generated:	December 08, 2025 at 01:30 AM
Department:	College of Engineering
Training Date:	December 05, 2025
Raw Individual Records:	364
Total Aggregated Records:	42
Best Performing Model:	Linear Regression
Model Accuracy (R^2):	1.000000
Number of Features:	11

TABLE OF CONTENTS

1. Introduction

This report documents the complete machine learning training process for the College of Engineering Board Exam Prediction System. The system uses historical anonymous board exam data to predict future passing rates using advanced regression algorithms. This AI-powered prediction system aims to help the institution make data-driven decisions regarding board exam preparation and student support programs.

The Engineering department covers various licensure examinations including Electronics Engineer (ECELE), Electronics Technician (ECTLE), Registered Electrical Engineer (REELE), and Registered Master Electrician (RMELE).

2. Data Collection

Data Source: The training data was collected from the LSPU Board Exam Records Management System, specifically from the *anonymous_board_passers* table in the MySQL database.

Department Filter: Only records from the "Engineering" department were included.

Collection Method: SQL query aggregating exam results by board exam type, year, month, and attempt type.

Data Period: 2021 to 2024

Raw Individual Records: 364 student examination records

Aggregated Records: 42 statistical records after grouping

Exam Types Covered:

• Electronics Engineer Licensure Examination (ECELE)

• Electronics Technician Licensure Exam (ECTLE)

• Registered Electrical Engineer Licensure Exam (REELE)

• Registered Master Electrician Licensure Exam (RMELE)

3. Data Cleaning and Preparation

The following data cleaning and preparation steps were performed:

a) Filtering Invalid Records:
• Excluded soft-deleted records (is_deleted = 1)
• Filtered only records from the Engineering department
• Removed records with null board_exam_date

b) Aggregation:
• Grouped data by board_exam_type, exam_year, exam_month, and attempt_type
• Calculated total_takers, total_passers, total_failed, and passing_rate for each group

c) Missing Value Handling:
• Records with null board_exam_date were excluded
• Passing rates calculated as (total_passers / total_takers) × 100
• Zero division handled with default values

d) Feature Engineering:
• Created year_normalized feature for temporal analysis (0-1 scale)
• Generated total_examinees count
• Computed first_timer_ratio and repeater_ratio binary indicators
• Calculated fail_rate and conditional_rate percentages
• Created passing_rate_ma3 (3-period moving average)
• One-hot encoded categorical exam types

4. Dataset Splitting (80% Training, 20% Testing)

The dataset was split into training and testing sets to ensure proper model validation:

Split Ratio: 80% Training / 20% Testing

Total Aggregated Records: 42
Training Set Size: 33 records (80%)
Testing Set Size: 9 records (20%)

Split Method: train_test_split from scikit-learn with random_state=42 for reproducibility

Purpose:
• **Training Set:** Used to train the machine learning models, allowing them to learn patterns from historical data
• **Testing Set:** Used to evaluate model performance on unseen data, ensuring the model generalizes well

Why 80-20 Split?
This is a standard split ratio that provides sufficient data for training while maintaining an adequate test set for reliable performance evaluation.

Dataset	Records	Percentage	Purpose
Training Set	33	80%	Model Learning
Testing Set	9	20%	Model Evaluation
Total	**42**	**100%**	**-**

5. Feature Selection

Feature selection identifies the most important variables that influence the prediction of passing rates. A total of **11 features** were selected based on their relevance to board exam performance:

Selected Features:

Feature Name	Description
year_normalized	Year converted to 0-1 scale for trend analysis
total_examinees	Number of students taking the exam
first_timer_ratio	Binary indicator (1 if first-time taker, 0 otherwise)
repeater_ratio	Binary indicator (1 if repeater, 0 otherwise)
fail_rate	Historical failure rate percentage
conditional_rate	Conditional passing rate percentage
passing_rate_ma3	3-period moving average of passing rates
exam_Electronics Engineer Licensure E...	Binary indicator for ECELE exam
exam_Electronics Technician Licensure...	Binary indicator for ECTLE exam

exam_Registered Electrical Engineer L...	Binary indicator for REELE exam
exam_Registered Master Electrician Li...	Binary indicator for RMELE exam

6. Model Selection

Seven different regression algorithms were selected and evaluated to find the best performing model for predicting board exam passing rates. These models represent a diverse set of approaches from simple linear methods to complex ensemble techniques:

Model	Type	Description
Linear Regression	Linear	Basic regression assuming linear relationship between features and target
Ridge Regression	Linear (L2)	Linear regression with L2 regularization to prevent overfitting
Lasso Regression	Linear (L1)	Linear regression with L1 regularization for feature selection
Random Forest	Ensemble	Ensemble of decision trees using bagging for improved accuracy
Gradient Boosting	Ensemble	Sequential ensemble method that corrects errors iteratively
XGBoost	Ensemble	Optimized gradient boosting with regularization
Support Vector Regression	Kernel-based	Finds optimal hyperplane for regression with RBF kernel

7. Model Training

The model training process was conducted as follows:

a) Data Preprocessing:
• Features scaled using StandardScaler (zero mean, unit variance)
• Categorical variables one-hot encoded
• Missing values handled appropriately

b) Training Process:
• Training Date: December 05, 2025
• Training Duration: Approximately 2-5 seconds per model
• All 7 models trained on the same training set (33 records)
• 5-Fold Cross-validation performed for robust evaluation

c) Hyperparameters Used:
• Random Forest: n_estimators=100, random_state=42
• Gradient Boosting: n_estimators=100, learning_rate=0.1
• XGBoost: n_estimators=100, learning_rate=0.1, max_depth=6
• Ridge/Lasso: alpha=1.0 (default regularization)
• SVR: kernel='rbf', C=1.0

d) Training Environment:
• Python 3.10+ with scikit-learn 1.x
• XGBoost 2.x for gradient boosting
• Models saved using joblib for persistence

8. Model Testing and Evaluation

After training, each model was evaluated on the held-out test set (9 records, 20% of data):

Evaluation Process:
• Models predict passing rates on test set
• Predictions compared to actual values
• Multiple metrics calculated for comprehensive evaluation
• Best model selected based on R^2 score and overall accuracy

Cross-Validation:
• 5-Fold cross-validation performed on training data
• Provides robust estimate of model performance
• Helps detect overfitting

Backtesting Validation:
• Additional validation by training on historical data
• Predicting known years to verify accuracy
• Comparing predicted vs actual values

9. Evaluation Metrics

The following metrics were used to evaluate model performance. These are standard metrics for regression problems:

Metric	Formula / Description	Interpretation
R^2 (R-Squared)	$R^2 = 1 - (SS_{res} / SS_{tot})$	Proportion of variance explained. Range: 0-1, higher is better. 1.0 = perfect fi
MAE (Mean Absolute Error)	$MAE = (1/n) \times \Sigma\|actual - predicted\|$	Average absolute difference. Lower is better. In percentage points.
MSE (Mean Squared Error)	$MSE = (1/n) \times \Sigma(actual - predicted)^2$	Average squared difference. Penalizes large errors more heavily.
RMSE (Root MSE)	$RMSE = \sqrt{MSE}$	Square root of MSE. Same unit as target variable (percentage).

Best Model Performance (Linear Regression):

Metric	Value	Notes
R^2 (R-Squared)	1.000000	Excellent fit
MAE (Mean Absolute Error)	0.000615%	Average error of 0.0006 percentage points
MSE (Mean Squared Error)	0.0000005158	Squared error metric
RMSE (Root MSE)	0.000718%	Typical error of ±0.0007%
Dataset Used	Engineering (42 records)	Years: 2021-2024

10. Prediction Generation

The prediction generation process works as follows:

a) Data Preparation:
• Fetch latest available data from database
• Prepare features using the same preprocessing pipeline
• Create next-year features based on latest data
• Apply the same StandardScaler transformation

b) Prediction Process:
• Load the best trained model (Linear Regression)
• Load the fitted StandardScaler from training
• Transform input features using the scaler
• Generate prediction using model.predict()

c) Output Generated:
• Predicted passing rate (0-100%)
• Prediction year (next year)
• Model used for prediction
• 95% Confidence interval bounds

d) Confidence Intervals:
• Calculated using historical prediction accuracy
• Provides upper and lower bounds for the prediction
• Helps quantify uncertainty in predictions

11. Complete Training Dataset

Total Records: 42

Data Period: 2021 to 2024

Board Exam Type	Records	Avg Passing Rate	Total Takers
Electronics Engineer Licensure E...	11	41.83%	77
Electronics Technician Licensure...	7	67.29%	56
Registered Electrical Engineer L...	12	41.17%	185
Registered Master Electrician Li...	12	44.15%	46
TOTAL	**42**	**46.55%**	**364**

12. Model Performance Comparison

All seven models were evaluated using the test set. The following table shows the complete comparison of model performance:

Model	Test R^2	Test MAE	Test MSE	CV Mean	CV Std
Linear Regression	1.000000	0.0006	0.000001	1.000000	0.000000
Lasso Regression	0.999986	0.0972	0.014059	0.999926	0.000075
Ridge Regression	0.997176	1.4157	2.891407	0.987917	0.007869
Random Forest	0.985701	2.8408	14.637487	0.911938	0.041495
Gradient Boosting	0.981779	2.1729	18.652353	0.924243	0.036502
XGBoost	0.971888	3.7989	28.778487	0.817958	0.089690
Support Vector Regression	-0.169169	28.0313	1196.870774	-0.643346	0.748043

Best Model Selected: Linear Regression (highlighted in green)

13. Conclusions and Recommendations

Key Findings:

1. **Model Performance:** The Linear Regression model achieved an R^2 score of 1.000000, indicating excellent predictive capability with near-perfect fit on the test data.

2. **Prediction Accuracy:** With an MAE of 0.000615%, the model's predictions are highly accurate, with typical errors less than 1 percentage point.

3. **Data Quality:** The aggregated dataset of 42 records from 2021-2024 provides sufficient statistical power for reliable predictions.

4. **Feature Importance:** Key predictive features include temporal trends (year_normalized), historical performance (passing_rate_ma3), and exam type indicators.

Recommendations:

1. **Regular Updates:** Retrain the model annually with latest exam results to maintain accuracy.

2. **Monitoring:** Track prediction accuracy against actual results for continuous validation.

3. **Data Collection:** Continue collecting comprehensive exam data to improve future predictions.

4. **Confidence Intervals:** Use the provided 95% confidence intervals when making decisions based on predictions.