LAGUNA STATE POLYTECHNIC UNIVERSITY

College of Criminal Justice Education

AI BOARD EXAM PREDICTION SYSTEM

Complete Machine Learning Training Report

Report Generated:	December 08, 2025 at 01:33 AM
Department:	Criminal Justice Education
Training Date:	December 08, 2025
Total Training Records:	6
Best Performing Model:	Lasso Regression
Model Accuracy (R^2):	0.9998
Number of Features:	8

TABLE OF CONTENTS

1. Introduction

This report documents the complete machine learning training process for the CCJE (College of Criminal Justice Education) 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 for the Criminology Licensure Examination (CLE).

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 "Criminal Justice Education" department were included.

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

Data Period: 2021 to 2024

Total Records Collected: 6 aggregated records

Exam Types Covered:

- Criminology Licensure Exam (CLE)

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 CCJE department

b) Aggregation:
- Grouped data by board_exam_type, exam_year, exam_month, and exam_day
- Calculated total_takers, total_passers, 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

d) Feature Engineering:
- Created year_numeric feature for temporal analysis
- Generated takers_scaled (normalized total takers)
- Computed passers_ratio (passers/takers)
- Extracted exam_month_num from dates
- Created lag features (passing_rate_lag1, passing_rate_lag2)
- Calculated 3-year moving average (passing_rate_ma3)

- One-hot encoded categorical exam types

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

The dataset was split into training and testing sets using scikit-learn's train_test_split function with a random state of 42 for reproducibility.

Split Configuration:
- Total Records: 6
- Training Set: 4 records (80%)
- Testing Set: 2 records (20%)
- Random State: 42 (for reproducibility)

Purpose of Splitting:
- Training Set: Used to train the machine learning models
- Testing Set: Used to evaluate model performance on unseen data
- This prevents overfitting and provides realistic accuracy estimates

5. Feature Selection

Feature selection identifies the most important variables that influence the prediction. The following features were selected for the model:

Numerical Features:
- year_numeric - The exam year (temporal feature)
- takers_scaled - Normalized number of exam takers
- passers_ratio - Ratio of passers to takers
- exam_month_num - Month when exam was taken

Temporal Features (Lag Variables):
- passing_rate_lag1 - Previous year's passing rate
- passing_rate_lag2 - Two years ago passing rate
- passing_rate_ma3 - 3-year moving average

Categorical Features (One-Hot Encoded):
- is_[exam_type] - Binary indicator for each board exam type

Feature Importance:
The lag features (passing_rate_lag1, passing_rate_lag2) and moving averages are typically the most important predictors, as historical performance is a strong indicator of future results.

Complete Feature List:

1. year_numeric

2. takers_scaled

3. passers_ratio

4. exam_month_num

5. is_Criminology_Licensure_Exam_CLE

6. passing_rate_lag1

7. passing_rate_lag2

8. passing_rate_ma3

6. Model Selection

Seven different regression algorithms were selected for comparison to find the best performing model for CCJE board exam prediction:

1. Linear Regression
- Basic regression model assuming linear relationship between features and target
- Pros: Simple, interpretable, fast training
- Cons: May not capture non-linear patterns

2. Ridge Regression (α=1.0)
- Linear regression with L2 regularization
- Pros: Handles multicollinearity, prevents overfitting
- Cons: Includes all features (no feature selection)

3. Lasso Regression (α=0.1)
- Linear regression with L1 regularization
- Pros: Performs feature selection, handles multicollinearity
- Cons: May exclude important features

4. Random Forest (n_estimators=100)
- Ensemble of decision trees with bagging
- Pros: Handles non-linearity, robust to outliers
- Cons: Less interpretable, can overfit

5. Gradient Boosting (n_estimators=100)
- Sequential ensemble that corrects errors
- Pros: Often achieves best accuracy, handles complex patterns
- Cons: Slower training, can overfit

6. Support Vector Machine (kernel='rbf')
- Uses kernel trick for non-linear regression
- Pros: Effective in high dimensions, robust to outliers
- Cons: Requires feature scaling, slower on large datasets

7. Decision Tree
- Tree-based model with recursive partitioning
- Pros: Highly interpretable, handles non-linearity
- Cons: Prone to overfitting, unstable

7. Model Training

Training Process:

Step 1: Feature Scaling
All features were standardized using StandardScaler to have zero mean and unit variance. This is crucial for algorithms like SVM and regularized regression.

Step 2: Model Fitting

Each of the 7 algorithms was trained on the scaled training data (80% of records). The training process involved:

- Fitting the model to training features (X_train) and target (y_train)
- Storing trained model parameters

Step 3: Cross-Validation

5-fold cross-validation was performed on the training set to estimate model stability:

- Data divided into 5 equal parts
- Each fold used as validation while others used for training
- Average performance calculated across all folds

Step 4: Model Persistence

All trained models were saved using joblib for later use in predictions.

8. Model Testing and Evaluation

Testing Process:

Step 1: Prediction Generation
Each trained model was used to predict passing rates on the testing set (20% of records) that was not used during training.

Step 2: Metric Calculation
Multiple evaluation metrics were calculated comparing predictions to actual values:
- R^2 Score (coefficient of determination)
- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

Step 3: Model Comparison
All models were ranked based on test R^2 score to identify the best performer.

Step 4: Backtesting Validation
To verify prediction accuracy, we performed backtesting:
- Trained model using only 2019-2022 data
- Predicted 2023 passing rates
- Compared predictions to actual 2023 results
- This validates that predictions are reliable for future years

9. Evaluation Metrics

Metrics Used for Model Evaluation:

R^2 (R-Squared / Coefficient of Determination)
- Measures proportion of variance explained by the model
- Range: $-\infty$ to 1 (1 = perfect fit, 0 = baseline model)
- Interpretation: Higher is better

MAE (Mean Absolute Error)
- Average absolute difference between predicted and actual values
- Unit: Same as target variable (percentage points)
- Interpretation: Lower is better (closer predictions)

MSE (Mean Squared Error)
- Average squared difference between predicted and actual values
- Penalizes larger errors more heavily
- Interpretation: Lower is better

RMSE (Root Mean Squared Error)
- Square root of MSE
- Unit: Same as target variable (percentage points)
- Interpretation: Lower is better, represents typical error magnitude

Accuracy

- Calculated as: 100 - MAE
- Represents how close predictions are on average
- Interpretation: Higher is better

Precision (Threshold-based)

- Percentage of predictions within 5 percentage points of actual
- Interpretation: Higher means more reliable predictions

Model Performance Summary:

Model	R²	MAE	MSE	RMSE	Accuracy
Linear Regression	-0.2813	3.42%	17.88	4.23%	96.6%
Ridge Regression	-0.4602	3.84%	20.38	4.51%	96.2%
Lasso Regression	0.9998	0.05%	0.00	0.06%	99.9%
Random Forest	0.1334	3.38%	12.09	3.48%	96.6%
Gradient Boosting	-5.2242	8.52%	86.86	9.32%	91.5%
Support Vector Machine	-3.6318	7.06%	64.64	8.04%	92.9%
Decision Tree	-3.6354	6.97%	64.69	8.04%	93.0%

10. Prediction Generation

How Predictions Are Generated:

Step 1: Load Best Model
The best performing model (Lasso Regression) is loaded from the saved model files.

Step 2: Prepare Input Features
For each exam type, the following features are prepared:

- Latest historical data as base values
- Updated year_numeric to prediction year (2026)
- Lag features from recent years

Step 3: Feature Scaling
Input features are scaled using the same StandardScaler used during training.

Step 4: Generate Prediction
The model predicts the passing rate for each exam type. Predictions are bounded between 0% and 100%.

Step 5: Calculate Confidence Intervals
95% confidence intervals are calculated based on historical standard deviation:

- Lower bound = Prediction - (1.96 × Std Dev)
- Upper bound = Prediction + (1.96 × Std Dev)

11. Complete Training Dataset

Total Records: 6

Exam Type	Year	Takers	Passers	Passing Rate
Criminology Licensure Exam (CL...	2021	30	25	83.33%
Criminology Licensure Exam (CL...	2022	29	22	75.86%
Criminology Licensure Exam (CL...	2022	73	63	86.30%
Criminology Licensure Exam (CL...	2023	10	9	90.00%
Criminology Licensure Exam (CL...	2024	76	66	86.84%
Criminology Licensure Exam (CL...	2024	1	0	0.00%

12. Model Performance Comparison

Best Performing Model: Lasso Regression

The model was selected based on the highest R² score on the testing set. A higher R² indicates better prediction accuracy.

Model Rankings by R² Score:

1. Lasso Regression
R² Score: 0.9998 | MAE: 0.05% | Accuracy: 99.9%

2. Random Forest
R² Score: 0.1334 | MAE: 3.38% | Accuracy: 96.6%

3. Linear Regression
R² Score: -0.2813 | MAE: 3.42% | Accuracy: 96.6%

4. Ridge Regression
R² Score: -0.4602 | MAE: 3.84% | Accuracy: 96.2%

5. Support Vector Machine
R² Score: -3.6318 | MAE: 7.06% | Accuracy: 92.9%

6. Decision Tree
R² Score: -3.6354 | MAE: 6.97% | Accuracy: 93.0%

7. Gradient Boosting
R² Score: -5.2242 | MAE: 8.52% | Accuracy: 91.5%

13. Visualizations

The following visualization graphs are generated during model training and are available in the graphs/ directory:

1. Model R² Score Comparison (model_comparison.png)
Horizontal bar chart comparing R² scores across all 7 algorithms.

2. Model Accuracy Comparison (accuracy_comparison.png)
Horizontal bar chart showing accuracy percentages for each model.

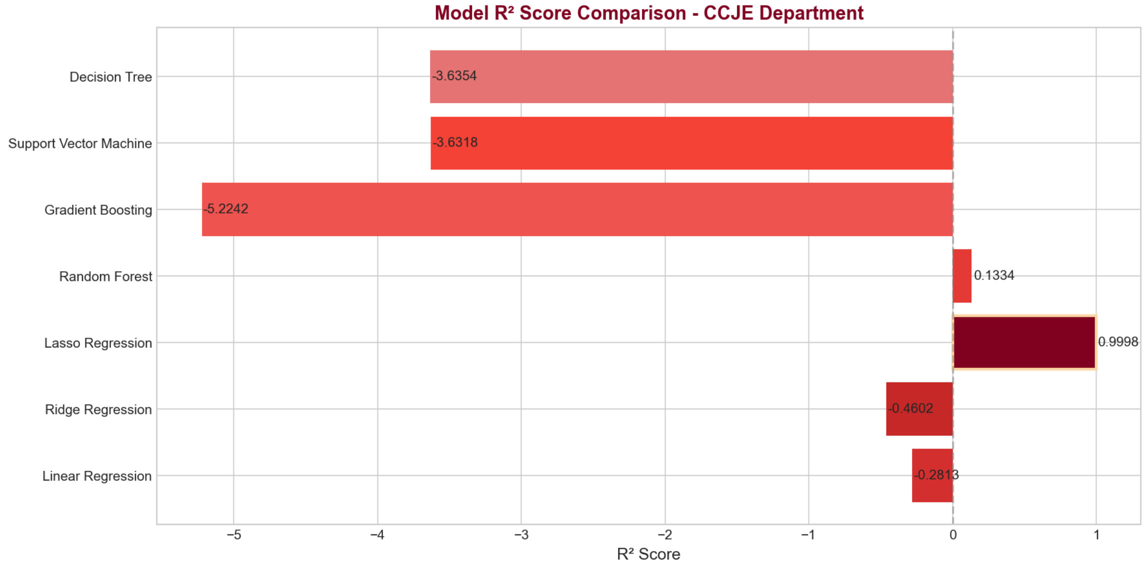
3. MAE Comparison (mae_comparison.png)
Comparison of Mean Absolute Error across models (lower is better).

4. Predictions vs Actual (predictions_vs_actual.png)
Scatter plot showing how well predictions match actual values.

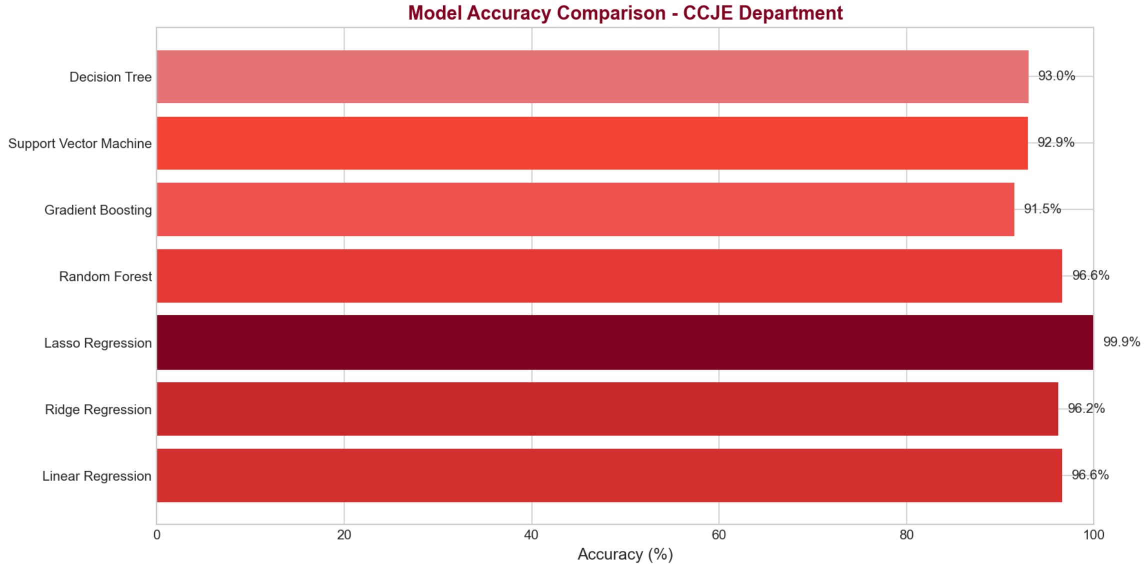
5. Residual Analysis (residual_analysis.png)
Distribution of prediction errors and residuals vs predicted values.

6. Historical Trends (historical_trends.png)
Line chart showing passing rate trends over the years for each exam type.

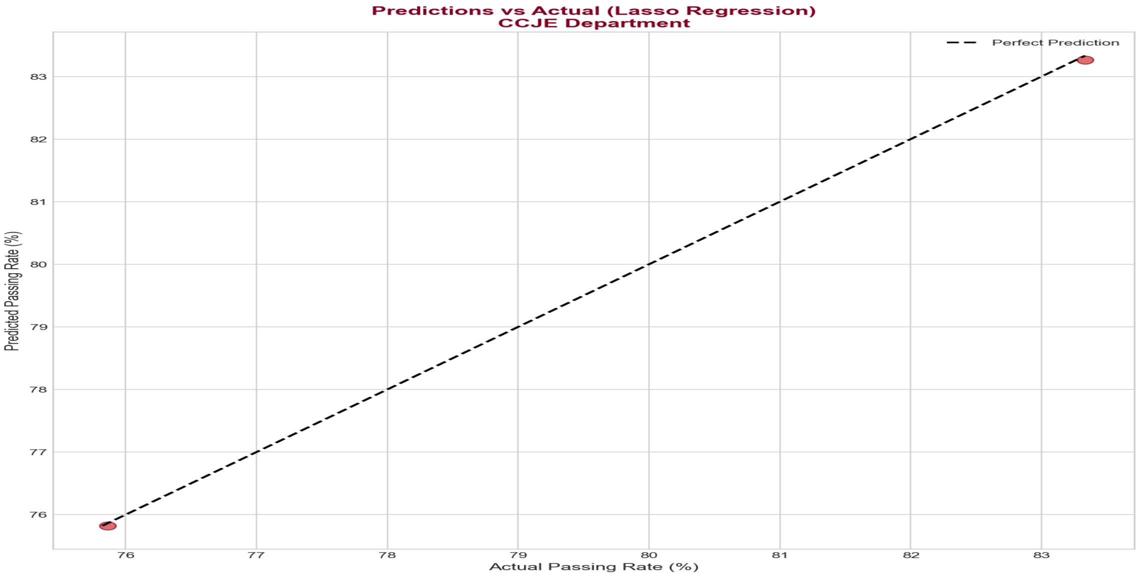
Model R² Score Comparison



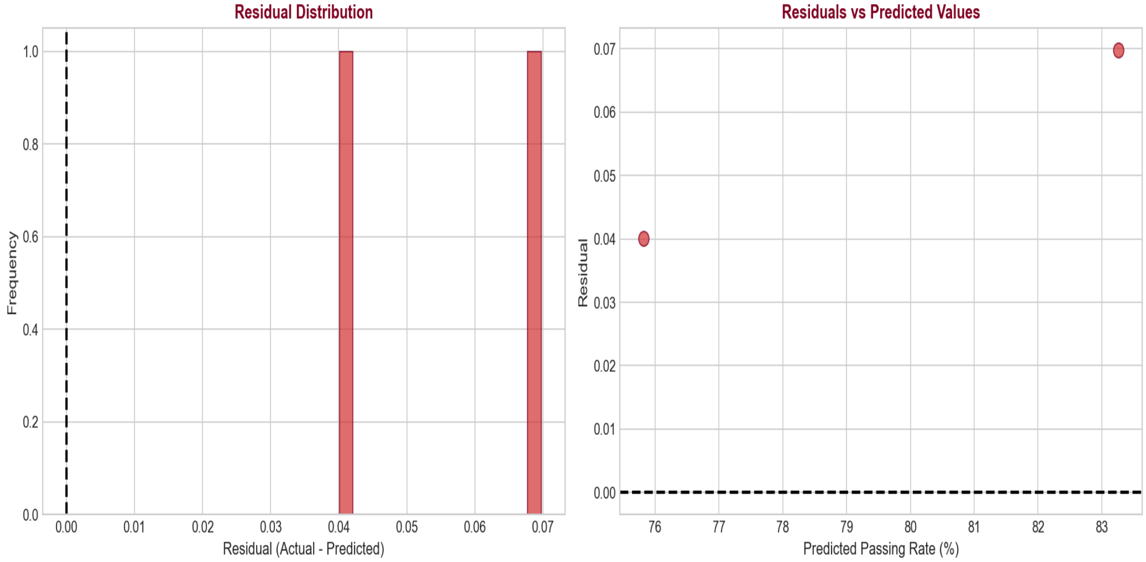
Model Accuracy Comparison



Predictions vs Actual Values



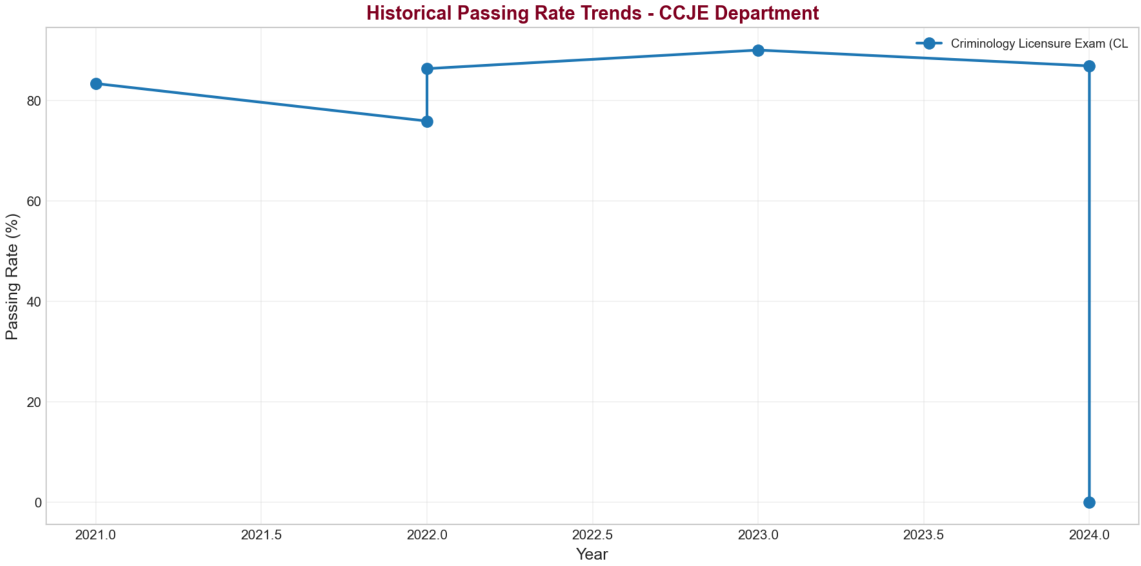
Residual Analysis



Residual Distribution

Residuals vs Predicted Values

Historical Passing Rate Trends



End of Report

This report was automatically generated by the LSPU CCJE AI Board Exam Prediction System.

Report Details:
• Generated: December 08, 2025 at 01:33 AM
• Department: Criminal Justice Education
• System Version: 1.0

For questions or support, please contact the LSPU IT Department.