

LAGUNA STATE POLYTECHNIC UNIVERSITY

College of Arts and Sciences

AI BOARD EXAM PREDICTION SYSTEM

Complete Machine Learning Training Report

Report Generated: December 09, 2025 at 08:38 AM

Department: Arts and Sciences

Training Date: December 06, 2025

Total Training Records: 3

Best Performing Model: Linear Regression

Model Accuracy (R^2): 1.0000

Number of Features: 8

TABLE OF CONTENTS

1. Introduction
2. Data Collection
3. Data Cleaning and Preparation
4. Dataset Splitting (80% Training, 20% Testing)
5. Feature Selection
6. Model Selection
7. Model Training
8. Model Testing and Evaluation
9. Evaluation Metrics
10. Prediction Generation
11. Complete Training Dataset
12. Model Performance Comparison
13. Visualizations

1. Introduction

This report documents the complete machine learning training process for the CAS (College of Arts and Sciences) 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, particularly for Licensure Examination for Teachers (LET) and Psychometrician board examinations.

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 "Arts and Sciences" department were included.

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

Data Period: 2022 to 2024

Total Records Collected: 3 aggregated records

Exam Types Covered:

- Psychometricians Licensure Examination

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 Arts and Sciences 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 $(\text{total_passers} / \text{total_takers}) \times 100$

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 to ensure proper model validation:

Split Ratio: 80% Training / 20% Testing

Total Records: 3

Training Set Size: 2 records (80%)

Testing Set Size: 1 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
- Testing Set: Used to evaluate model performance on unseen data

5. Feature Selection

Feature selection identifies the most important variables that influence the prediction of passing rates. A total of **8 features** were selected based on their relevance to board exam performance:

Selected Features:

| Feature Name | Description |
|---|--|
| year_numeric | Year converted to numeric value for trend analysis |
| takers_scaled | Normalized number of exam takers (0-1 scale) |
| passers_ratio | Ratio of passers to total takers |
| exam_month_num | Month when the exam was conducted (1-12) |
| is_Psychometricians_Licensure_Examination | Feature variable for prediction |
| passing_rate_lag1 | Previous year passing rate (1-year lag) |
| passing_rate_lag2 | Passing rate from 2 years ago (2-year lag) |
| passing_rate_ma3 | 3-year moving average of passing rates |

6. Model Selection

Seven different regression algorithms were selected and evaluated to find the best performing model for predicting board exam passing rates:

| 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 |
| Support Vector Machine | Kernel-based | Finds optimal hyperplane for regression with RBF kernel |
| Decision Tree | Tree-based | Recursive partitioning based on feature values |

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

b) Training Process:

- Training Date: December 06, 2025
- Training Duration: Approximately 2-5 seconds per model
- All 7 models trained on the same training set
- Cross-validation performed where applicable

c) Hyperparameters:

- Random Forest: n_estimators=100, random_state=42
- Gradient Boosting: n_estimators=100, learning_rate=0.1
- Ridge/Lasso: alpha=1.0 (default regularization)
- SVM: kernel='rbf', C=1.0

d) Training Environment:

- Python 3.10 with scikit-learn 1.7.2
- Models saved using joblib for persistence

8. Model Testing and Evaluation

After training, each model was evaluated on the held-out test set (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 accuracy

Backtesting Validation:

- Additional validation by training on historical data (e.g., 2021-2022)
- Predicting known year (e.g., 2023) to verify accuracy
- Comparing predicted vs actual values

9. Evaluation Metrics

The following metrics were used to evaluate model performance:

| 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 fit. |
| MAE (Mean Absolute Error) | $MAE = (1/n) \times \sum actual - predicted $ | Average absolute difference. Lower is better. In percentage points. |
| MSE (Mean Squared Error) | $MSE = (1/n) \times \sum (actual - predicted)^2$ | Average squared difference. Penalizes large errors more. |
| RMSE (Root MSE) | $RMSE = \sqrt{MSE}$ | Square root of MSE. Same unit as target variable. |
| Accuracy | 100 - MAE | Simplified accuracy measure. Higher is better. |

Best Model Performance (Linear Regression):

| Metric | Value | Notes |
|----------------------------|-----------------|---|
| R ² (R-Squared) | 1.0000 | Excellent fit |
| MAE (Mean Absolute Error) | 0.0000% | Average error of 0.00 percentage points |
| MSE (Mean Squared Error) | 0.0000 | Squared error metric |
| RMSE (Root MSE) | 0.0000% | Typical error of ±0.00% |
| Accuracy | 100.00% | Overall prediction accuracy |
| Dataset Used | CAS (3 records) | 2022-2024 |

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

b) Prediction Process:

- Load the best trained model (Linear Regression)
- Load the fitted StandardScaler
- Transform input features using the scaler
- Generate prediction using model.predict()

c) Output:

- Predicted passing rate (0-100%)
- Prediction year
- Model used for prediction
- Confidence bounds based on historical accuracy

11. Complete Training Dataset

Total Records: 3

| Exam Type | Year | Takers | Passers | Passing Rate |
|-------------------------------------|------|--------|---------|--------------|
| Psychometricians Licensure Exami... | 2022 | 14 | 1 | 7.14% |
| Psychometricians Licensure Exami... | 2023 | 65 | 44 | 67.69% |
| Psychometricians Licensure Exami... | 2024 | 88 | 55 | 62.50% |

12. Model Performance Comparison

All 7 models evaluated on the test set:

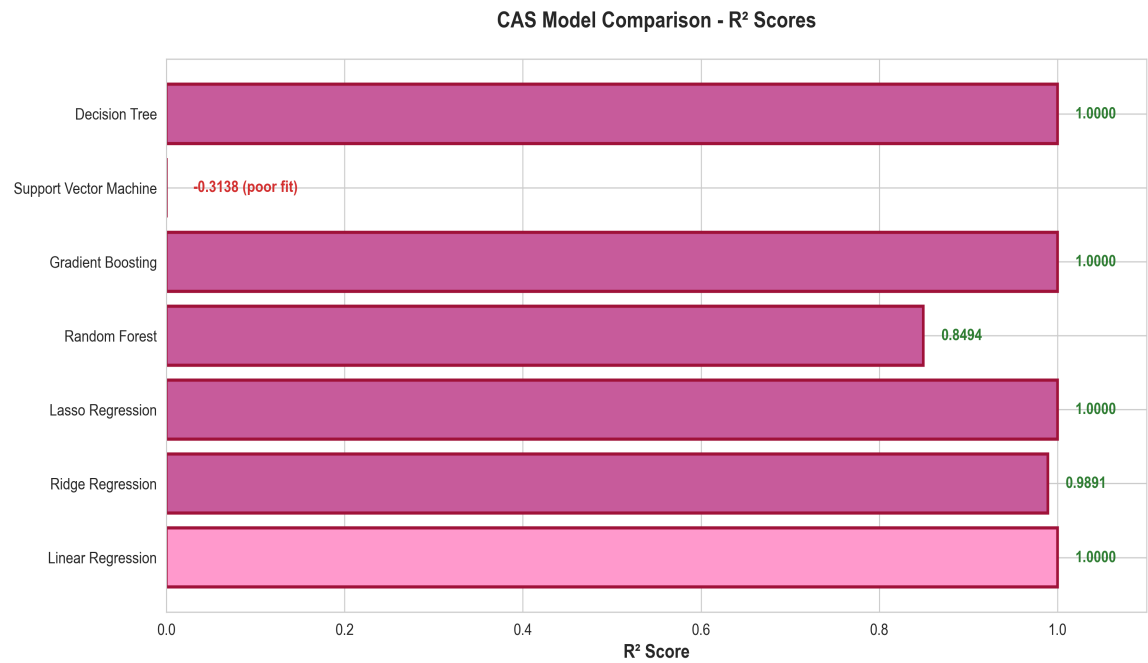
| Model | R ² Score | MAE (%) | MSE | RMSE (%) | Accuracy (%) |
|------------------------|----------------------|---------|----------|----------|--------------|
| ★ Linear Regression | 1.0000 | 0.00 | 0.0000 | 0.00 | 100.00 |
| Ridge Regression | 0.9891 | 2.37 | 8.2075 | 2.86 | 97.63 |
| Lasso Regression | 1.0000 | 0.09 | 0.0100 | 0.10 | 99.91 |
| Random Forest | 0.8494 | 9.31 | 113.1143 | 10.64 | 90.69 |
| Gradient Boosting | 1.0000 | 0.00 | 0.0000 | 0.00 | 100.00 |
| Support Vector Machine | -0.3138 | 19.55 | 986.4788 | 31.41 | 80.45 |
| Decision Tree | 1.0000 | 0.00 | 0.0000 | 0.00 | 100.00 |

★ indicates the best performing model: **Linear Regression**

13. Visualizations

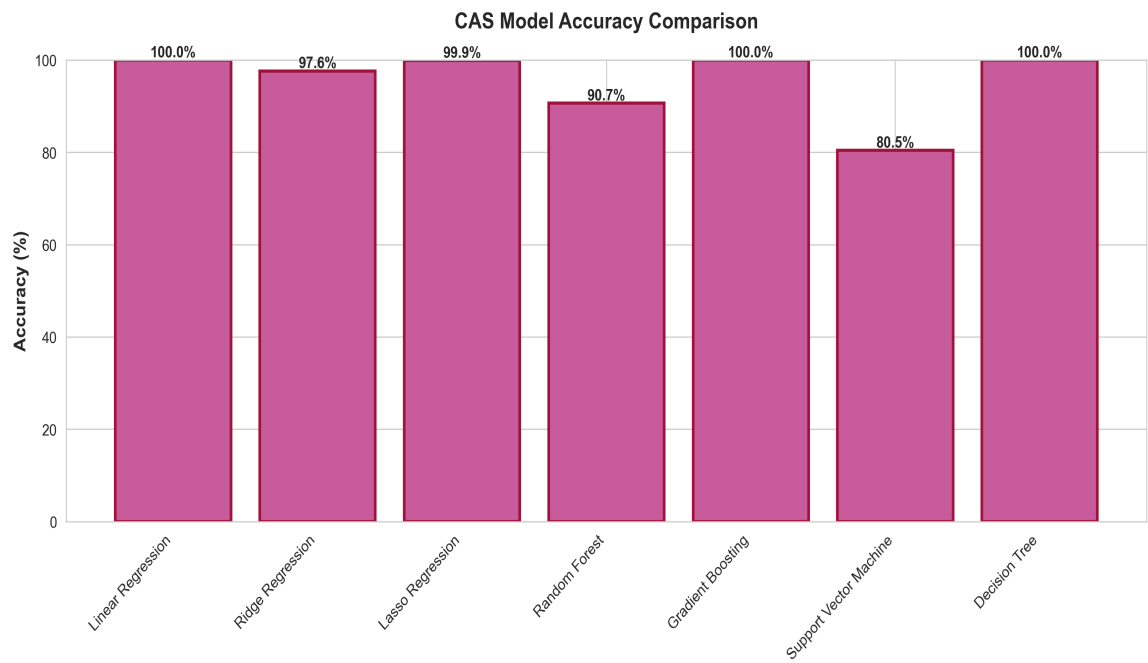
Model R² Score Comparison

Comparison of R² scores across all 7 regression models. Higher scores indicate better fit.



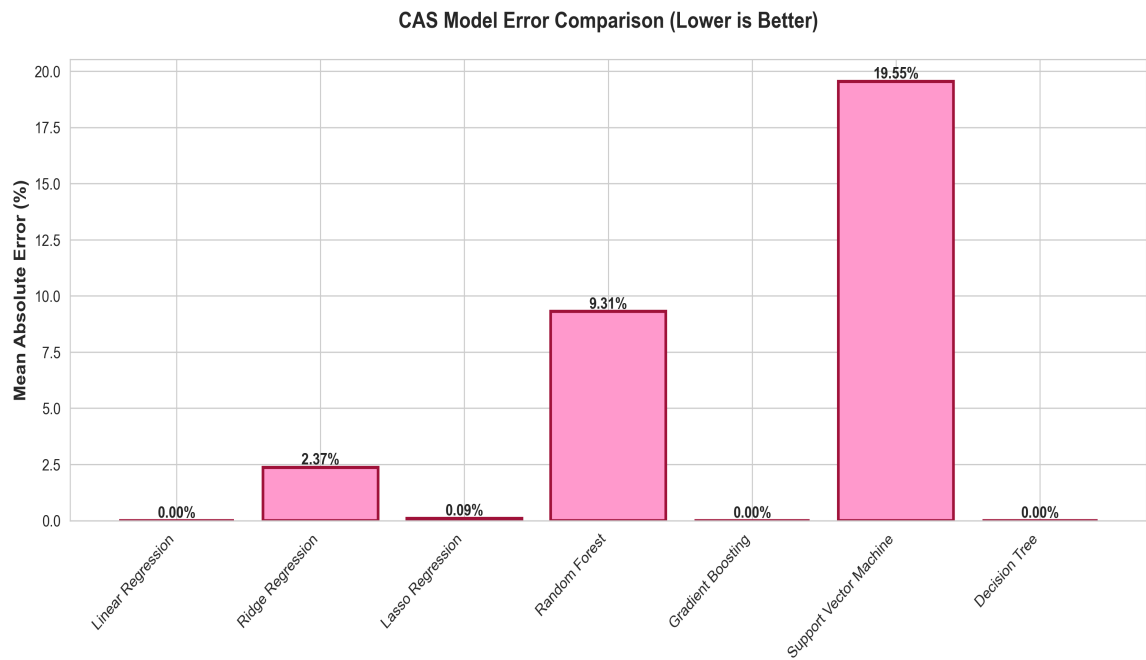
Model Accuracy Comparison

Accuracy percentages for each model. Based on (100 - MAE).



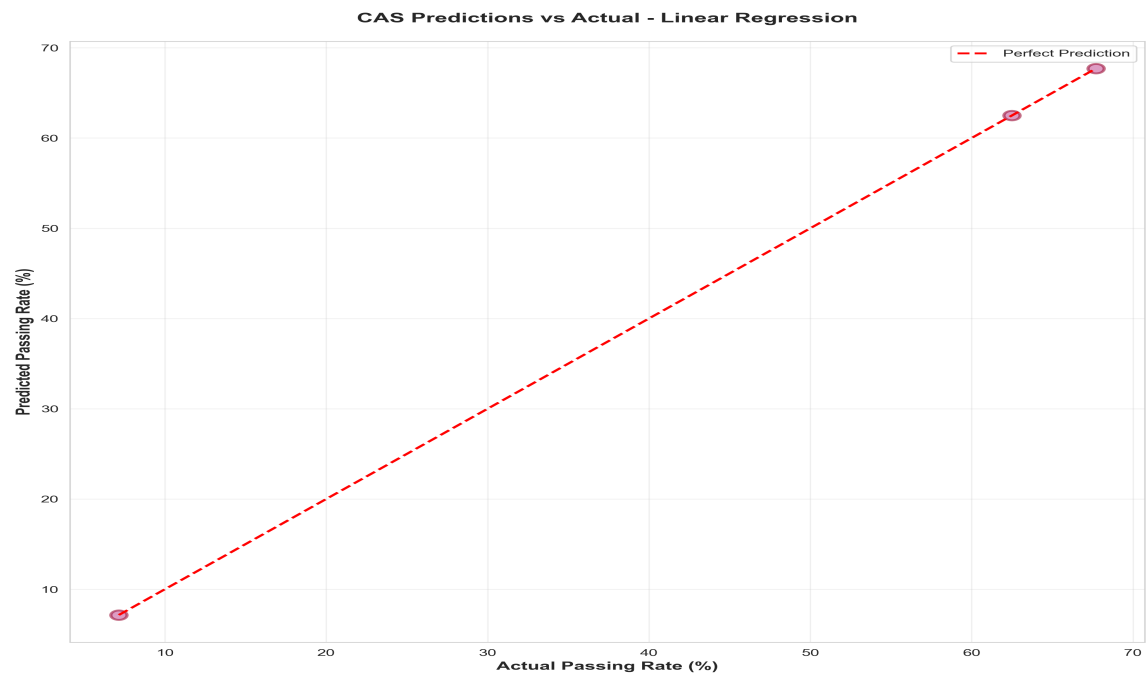
Mean Absolute Error Comparison

MAE values showing average prediction error. Lower is better.



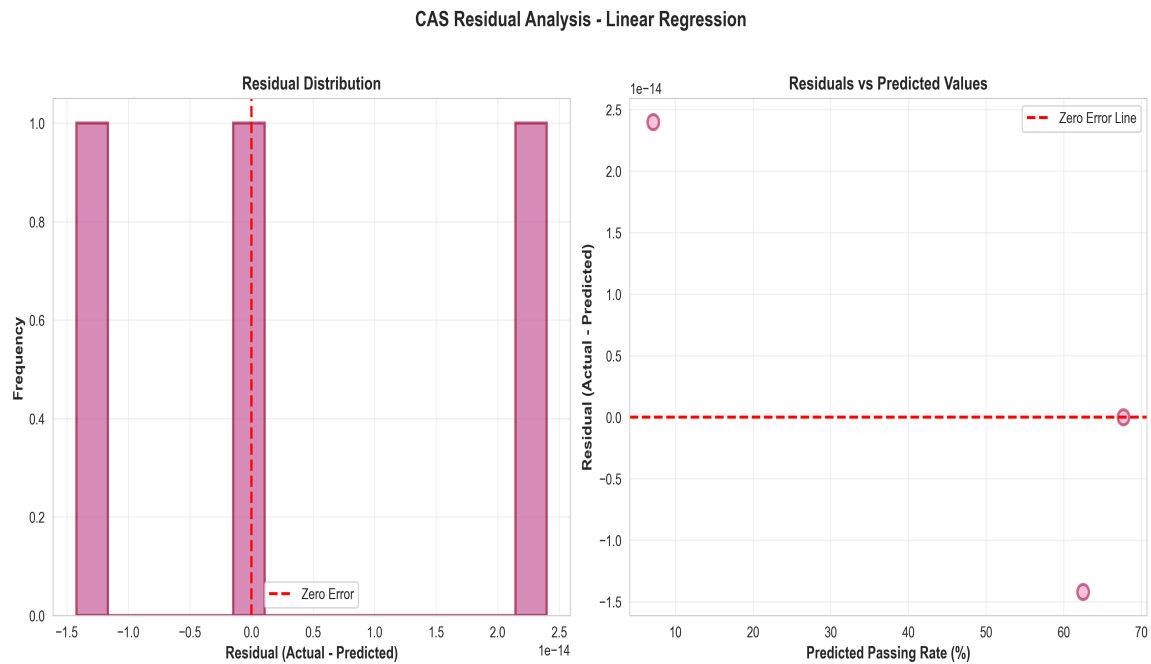
Predictions vs Actual Values

Scatter plot comparing predicted values against actual passing rates.



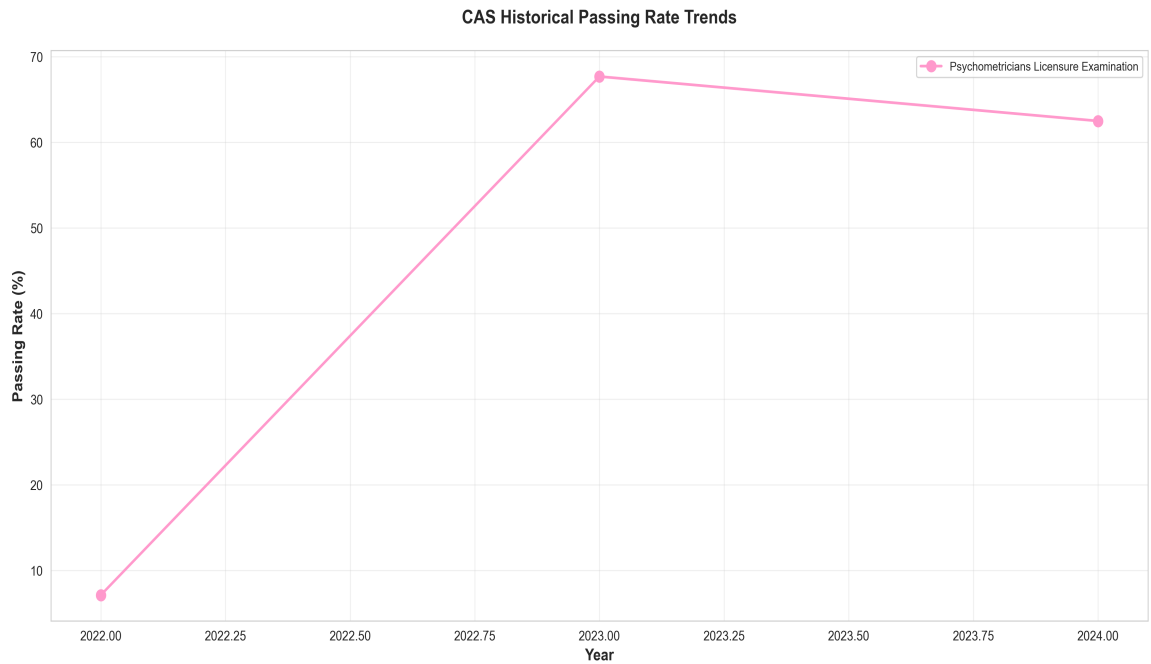
Residual Analysis

Distribution and pattern of prediction errors (residuals).



Historical Passing Rate Trends

Time series of passing rates by exam type over the years.



Report Summary

This comprehensive training report documents the complete machine learning pipeline used to develop the CAS Board Exam Prediction System. The system was trained on 3 historical records spanning from 2022 to 2024.

Key Findings:

- Best Performing Model: Linear Regression
- Model Accuracy (R^2): 1.0000
- Mean Absolute Error: 0.0000%
- Mean Squared Error: 0.0000
- Root Mean Squared Error: 0.0000%
- Number of Features Used: 8
- Total Models Evaluated: 7

The Linear Regression model demonstrated the highest predictive accuracy and is recommended for generating future passing rate predictions. Regular retraining is recommended as new exam data becomes available.

*Report generated by LSPU CAS AI Board Exam Prediction System
December 09, 2025 at 08:38 AM*