AutoJudge

Problem Difficulty Prediction System

Project Report - January 8, 2026

1. Executive Summary

AutoJudge is a machine learning-based system designed to automatically predict the difficulty level of competitive programming problems. The system analyzes problem descriptions, input/output specifications, and other textual features to classify problems into difficulty categories (Easy, Medium, Hard) and predict difficulty scores.

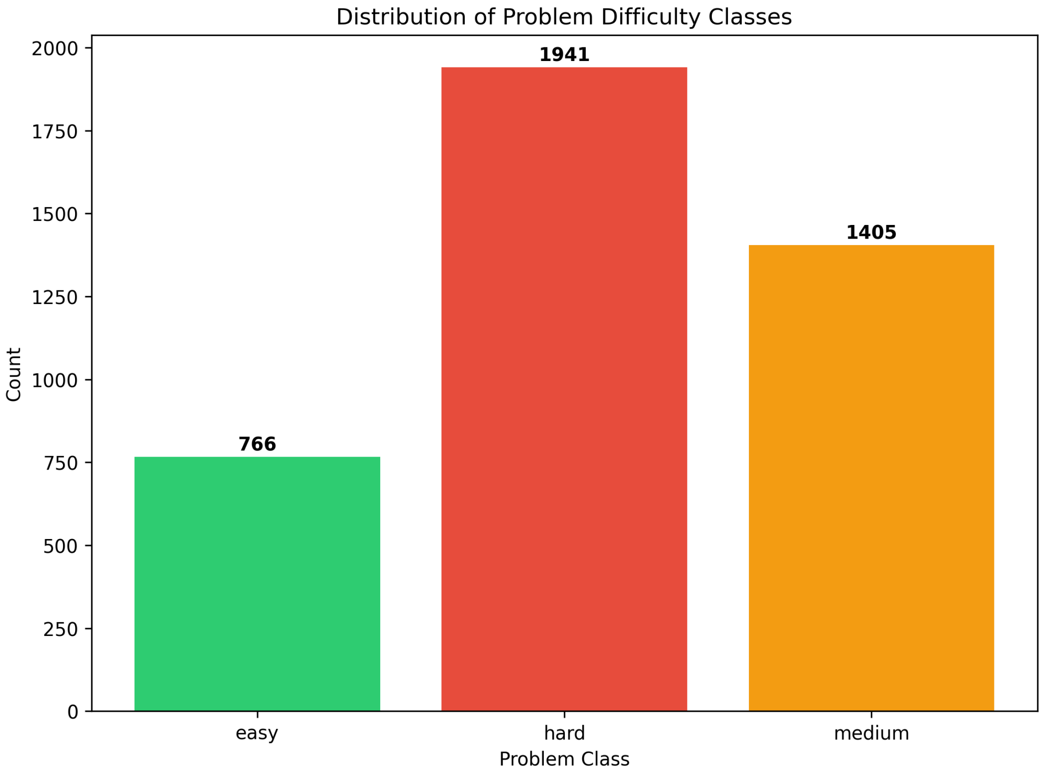
2. Dataset Overview

2.1 Dataset Statistics

Metric	Value
Total Samples	4,112
Training Set	3,289 (80%)
Test Set	823 (20%)
Total Features	1,094

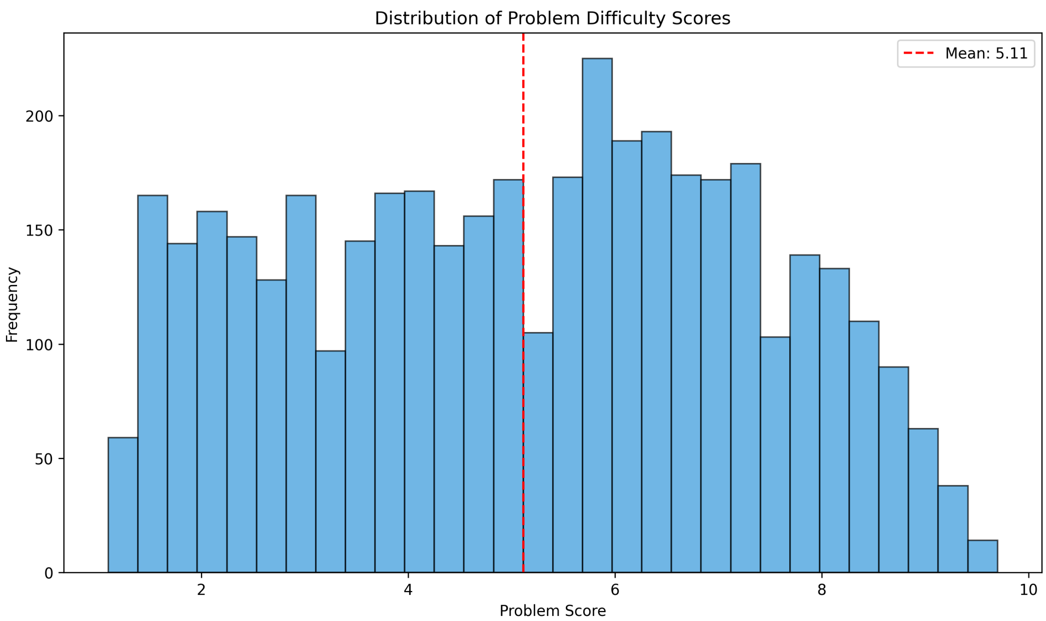
2.2 Class Distribution

Difficulty Class	Count	Percentage
Easy	766	18.6%
Medium	1,405	34.2%
Hard	1,941	47.2%



2.3 Score Distribution

The problem difficulty scores range from 1.10 to 9.70 on a 10-point scale.



3. Feature Engineering

3.1 Text Preprocessing

- Combined multiple text fields: title, description, input/output specifications

- Applied text cleaning and normalization

- Removed stop words and special characters

3.2 Feature Categories

Basic Statistical Features

Text length, word count, average word length, mathematical symbol count, digit patterns, power notation detection

Algorithm Keyword Features

Detection of 50+ algorithm-related keywords including: Graph algorithms (DFS, BFS, Dijkstra), Dynamic Programming, Data structures (Tree, Heap, Stack), Advanced concepts (Segment Tree, Trie, FFT)

TF-IDF Features

N-gram range: (1, 3), Maximum features: 1,000, Sublinear TF scaling applied

4. Model Architecture

4.1 Classification Model

Model Type: Random Forest Classifier

Parameter	Value
Number of Estimators	200
Maximum Depth	20
Min Samples Split	5
Min Samples Leaf	2

4.2 Regression Model

Model Type: Gradient Boosting Regressor (Best performing)

Parameter	Value
Number of Estimators	300
Learning Rate	0.05
Maximum Depth	6
Subsample Ratio	0.8

5. Classification Results

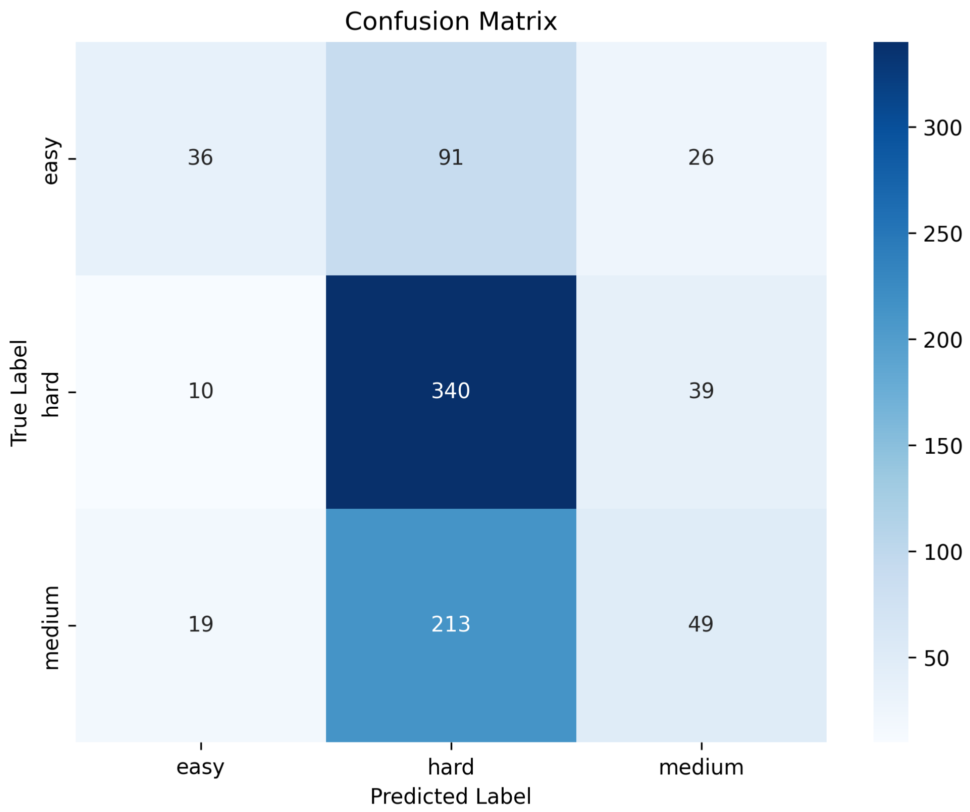
5.1 Overall Performance

Metric	Value
Accuracy	51.64%
Macro Avg F1-Score	0.41
Weighted Avg F1-Score	0.46

5.2 Per-Class Performance

Class	Precision	Recall	F1-Score
Easy	0.55	0.24	0.33
Medium	0.43	0.17	0.25
Hard	0.53	0.87	0.66

5.3 Confusion Matrix



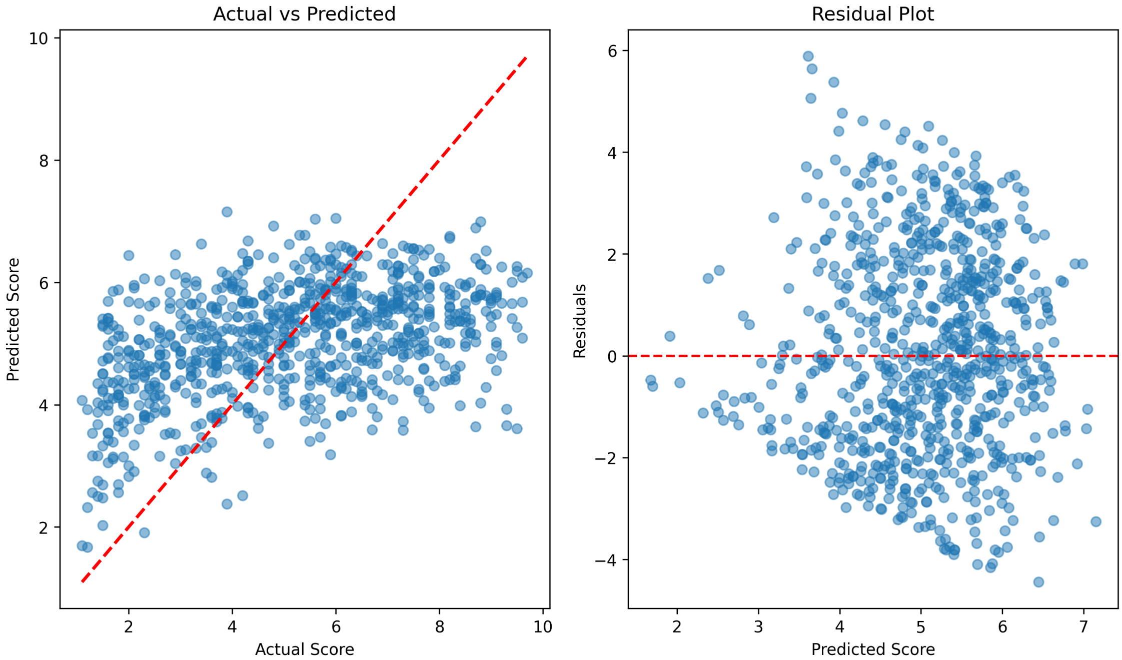
Analysis: The model shows strong performance in identifying "Hard" problems (87% recall). "Easy" and "Medium" classes show lower recall due to class imbalance.

6. Regression Results

6.1 Error Metrics

Metric	Value
Mean Absolute Error (MAE)	1.66
Root Mean Squared Error (RMSE)	2.00

6.2 Prediction Analysis



On average, predictions are within +/- 1.66 points of actual scores on a 10-point scale.

7. Models Compared

Model	RMSE	Performance
Gradient Boosting	1.99	Best
XGBoost	2.01	Good
Extra Trees	2.02	Moderate
Random Forest	2.04	Baseline

8. Files and Artifacts

8.1 Trained Models

File	Description
classifier.pkl	Trained classification model
classifier_scaler.pkl	Feature scaler for classifier
regressor.pkl	Trained regression model
regressor_scaler.pkl	Feature scaler for regressor
feature_extractor.pkl	Fitted feature extractor

9. Future Improvements

1. Data Augmentation: Collect more samples for underrepresented classes
2. Advanced NLP: Incorporate transformer-based embeddings (BERT, RoBERTa)
3. Ensemble Methods: Combine multiple models for better predictions
4. Feature Selection: Apply feature importance analysis
5. Active Learning: Continuously improve with user feedback

10. Conclusion

AutoJudge successfully demonstrates the feasibility of automated problem difficulty assessment using machine learning. The system achieves:

- 51.64% classification accuracy across three difficulty levels

- MAE of 1.66 for score prediction on a 10-point scale

- Strong identification of hard problems (87% recall)

The model provides a solid foundation for assisting competitive programming platforms in automatically categorizing problem difficulty, with clear paths for future enhancement.

Shivansh Yadav