AuToJudge Classifier

Automated Difficulty Estimation for Competitive Programming Problems

Bhavya Shah

January 2026

Abstract

AuToJudge Classifier is a machine learning system designed to automatically estimate the difficulty of competitive programming problems. The system predicts both a categorical difficulty label (Easy, Medium, Hard) and a numerical difficulty score on a 10-point scale. By combining semantic text embeddings with structural features extracted from problem statements and constraints, the system achieves a classification accuracy of 50.8% and a regression mean absolute error of 1.68 on the TaskComplexityEval-24 dataset. The complete system is deployed as a real-time web application with an interactive user interface.

1 Introduction

Competitive programming platforms host thousands of algorithmic problems with varying difficulty levels. Accurate difficulty estimation plays a key role in contest design, learning progression, and fair evaluation. However, manual difficulty labeling is subjective and often inconsistent.

This project proposes an automated system that analyzes problem descriptions, constraints, and structural patterns to estimate difficulty using machine learning techniques.

1.1 Objectives

1. Predict difficulty class (Easy, Medium, Hard)
2. Predict a continuous difficulty score
3. Combine semantic and structural indicators
4. Deploy a real-time usable system

2 Dataset

2.1 Source

The TaskComplexityEval-24 dataset consists of competitive programming problems collected from multiple platforms. After preprocessing, 3,899 problems were retained.

2.2 Class Distribution

- Easy: 18.6%
- Medium: 34.0%
- Hard: 47.3%

The dataset is highly imbalanced, with Hard problems forming the largest category.

3 Methodology

3.1 Text Preprocessing

Problem descriptions, input formats, and output formats were merged into a single text representation. Mathematical expressions were normalized, and unnecessary formatting was removed.

3.2 Feature Engineering

Two complementary feature sets were used. Semantic features capture the overall meaning of the problem using sentence embeddings. Structural features capture measurable indicators such as constraint size, algorithmic keywords, and input-output structure.

The final feature vector consists of 401 dimensions.

4 Models and Their Behavior

4.1 Baseline Models

Logistic Regression and Support Vector Machines were evaluated as baseline classifiers. These models struggled to separate Medium and Hard problems due to overlapping characteristics and the non-linear nature of difficulty.

4.2 LightGBM Classifier

LightGBM, a tree-based ensemble model, achieved the best classification performance. It effectively captured patterns such as large constraints and advanced algorithmic terminology, though some overfitting was observed.

Performance:

- Accuracy: 50.8%
- Macro F1-score: 0.46

4.3 Regression Models

Ridge Regression served as a baseline, while Gradient Boosting achieved the best balance between accuracy and generalization.

Regression performance:

- MAE: 1.68
- RMSE: 2.00

5 System Architecture

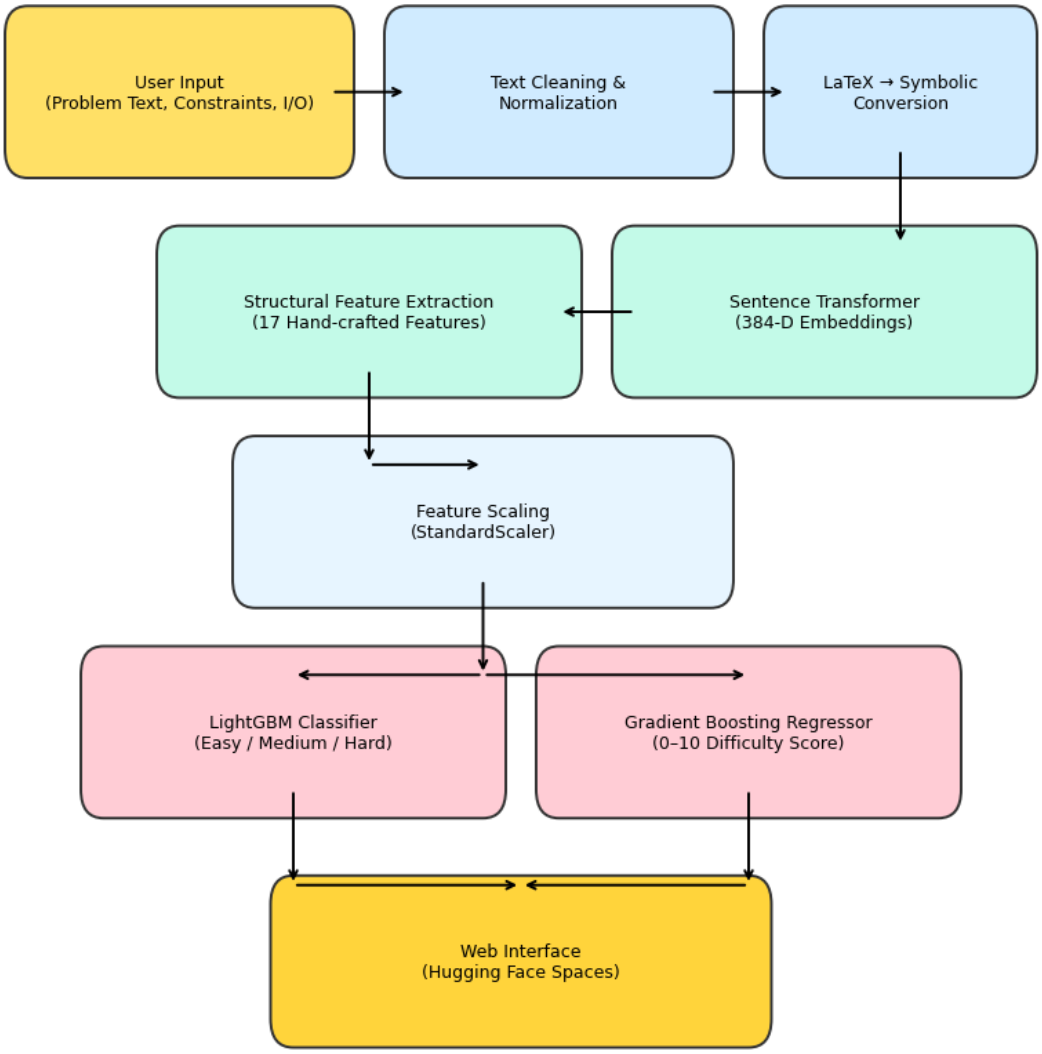


Figure 1: End-to-end architecture of the AuToJudge system

The system consists of a Flask backend, a custom feature extractor, and trained machine learning models used for prediction.

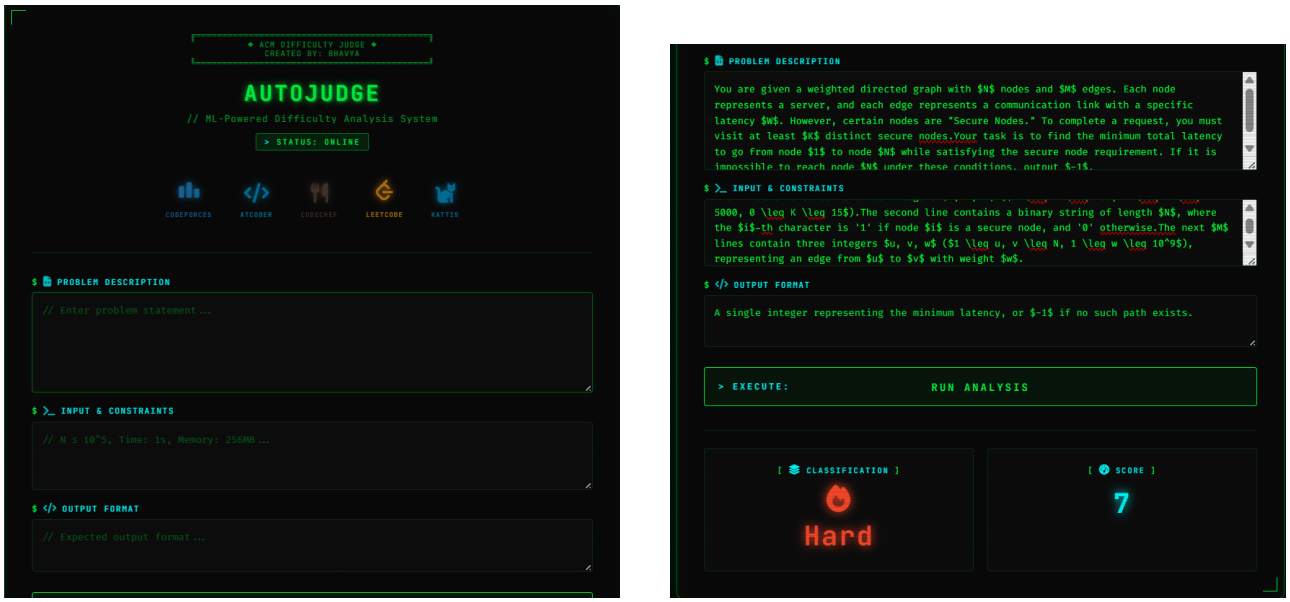
6 Web Interface

6.1 Frontend

The frontend is implemented using HTML, CSS, and JavaScript with a terminal-inspired dark theme.

6.2 Backend

The backend exposes a REST API that returns difficulty predictions in real time.



(a) Input interface

(b) Prediction output

Figure 2: AuToJudge web application

7 Results and Discussion

The classifier performs best on Hard problems due to dominant signals such as large constraints. Medium problems are frequently misclassified, reflecting ambiguity between difficulty levels. The regression model provides useful approximate scoring.

8 Conclusion

This project demonstrates that automated difficulty estimation for competitive programming problems is feasible. While perfect accuracy is unrealistic due to subjectivity, the system provides meaningful predictions that can assist students, educators, and contest organizers.