Analysis of Educational Data and Student Number Prediction Using WEKA

Co-funded by the Erasmus+ Programme of the European Union

Croatian team

ABOUT GIMNAZIJA METKOVIC

- Gimnazija Metkovic is a state secondary school with around 350 students age 14-19 and around 40 teachers.
- There are three main programmes - language-oriented, STEM-oriented and comprehensive.
- The curriculum is highly academic with around 15 compulsory subjects each year, plus several electives.
- All pupils go on to higher education at university. Our students have consistently scored high in numerous competitions and national final exams.

school projects - erasmus

- Gimnazija Metković has participated in Erasmus plus since 2014., and there have been numberous partnerships with schools all over Europe. We are also currently coordinating an Erasmus project, are partnering in several of them, and are an accredited school.

Project Objective

- The primary objectives of this study are to examine the relationships between key variables, such as the number of kindergartens, schools, and students, and to estimate future trends in student enrollment across various educational levels.
- This analysis aims to provide valuable insights into the factors influencing the growth and distribution of educational resources in Croatia.

dataset

HZZS – dataset

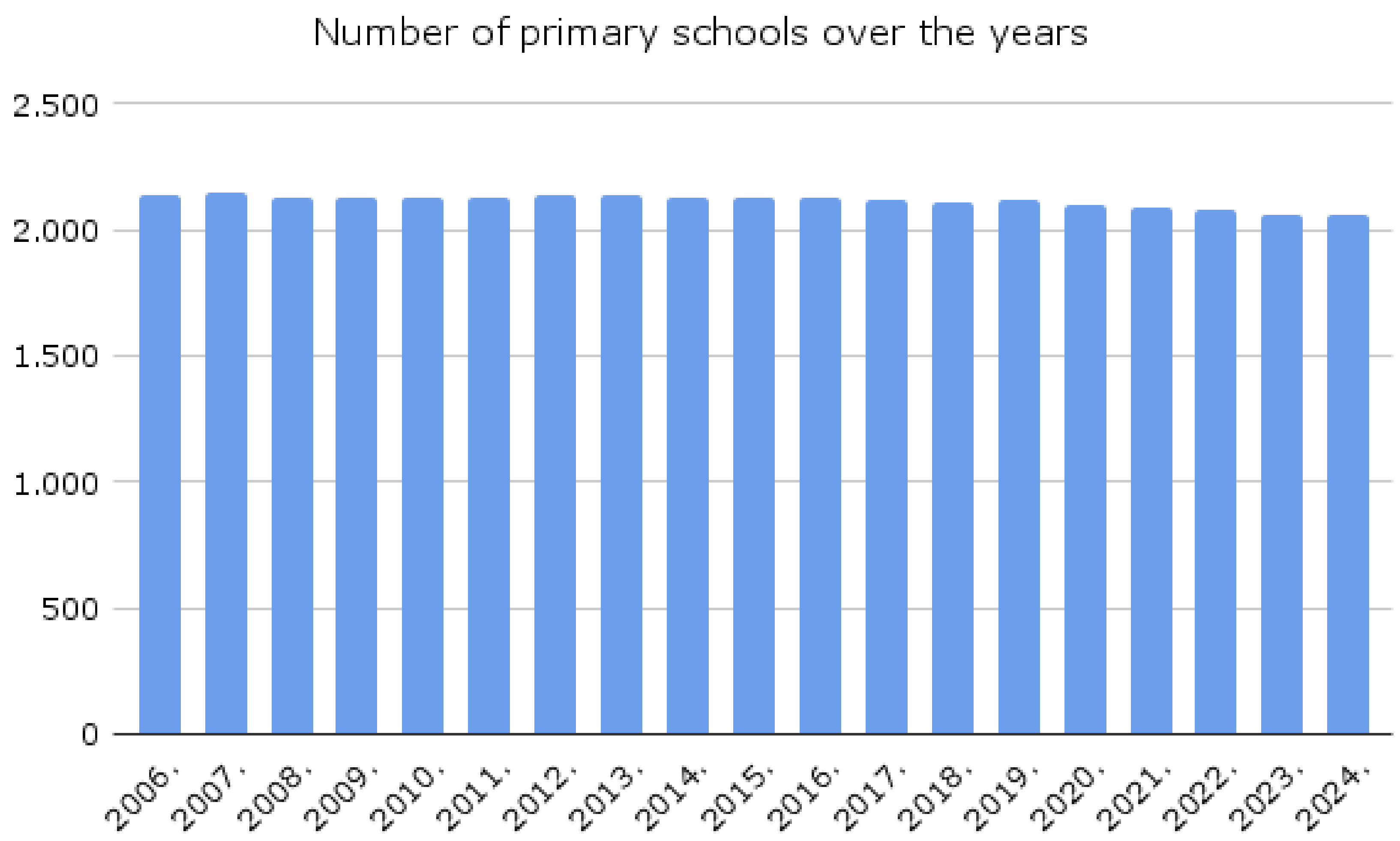
–number of schools, highschools and kindergartens in all Croatian countys

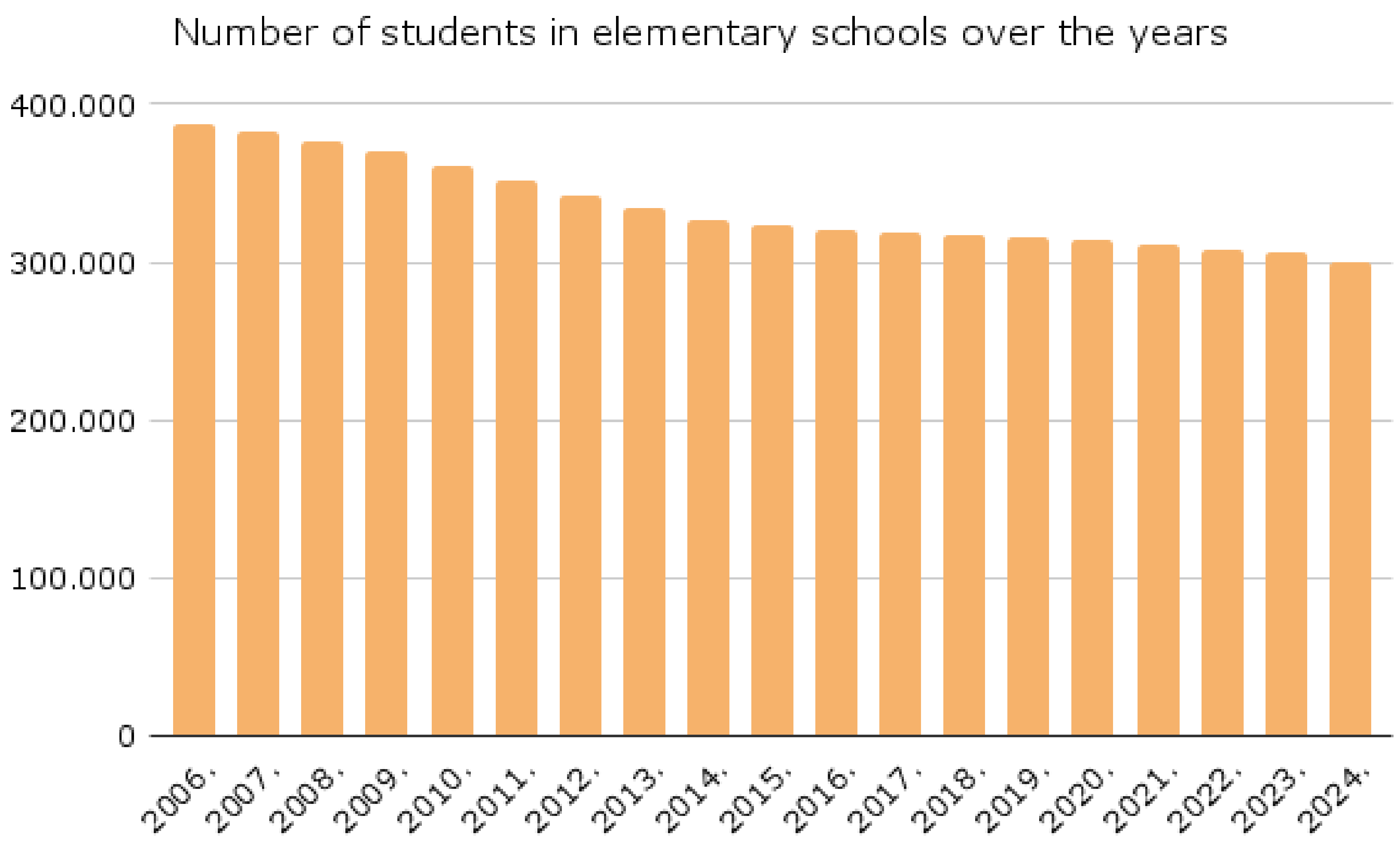
–2006–2024

predictions for the next year

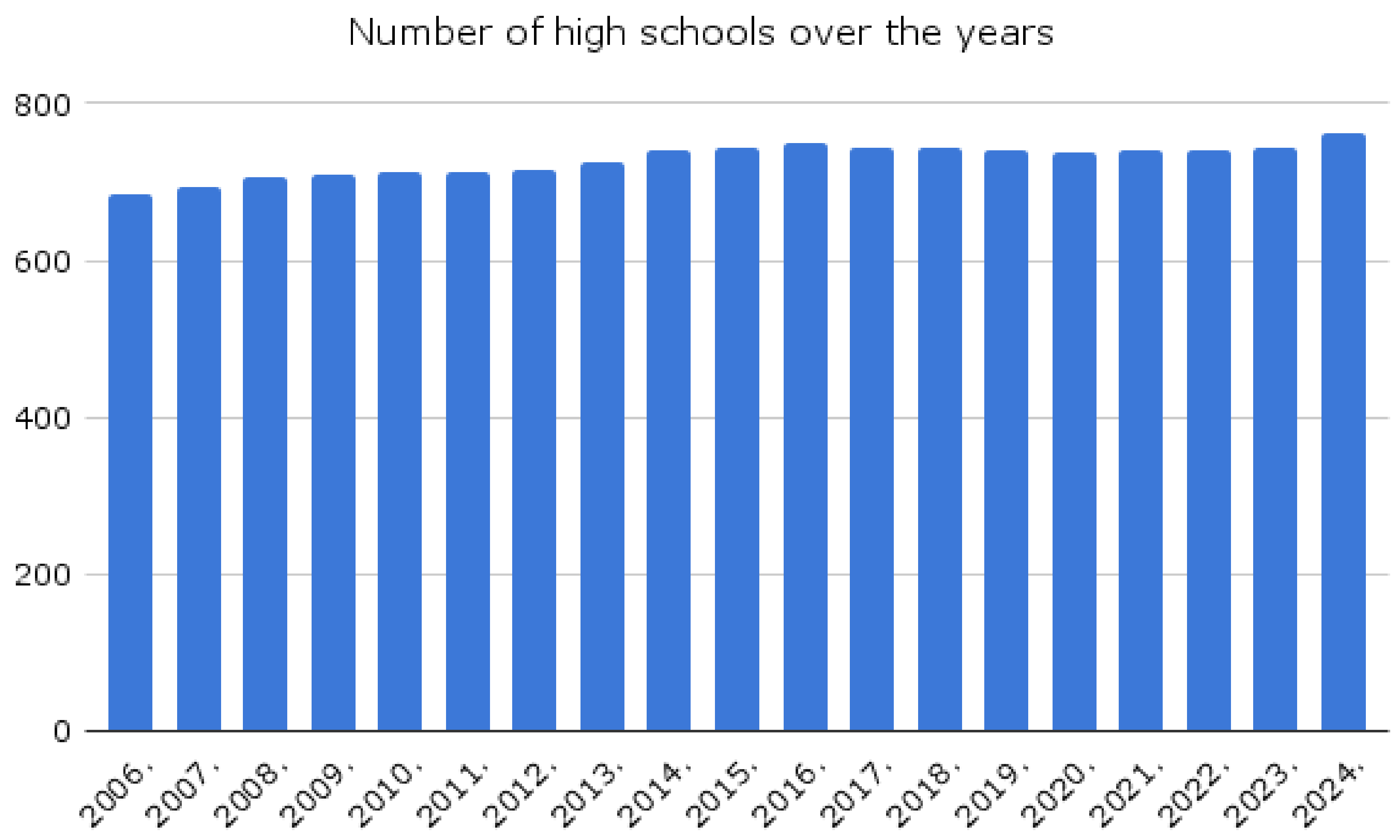
DRŽAVNI ZAVOD ZA STATISTIKU
CROATIAN BUREAU OF STATISTICS

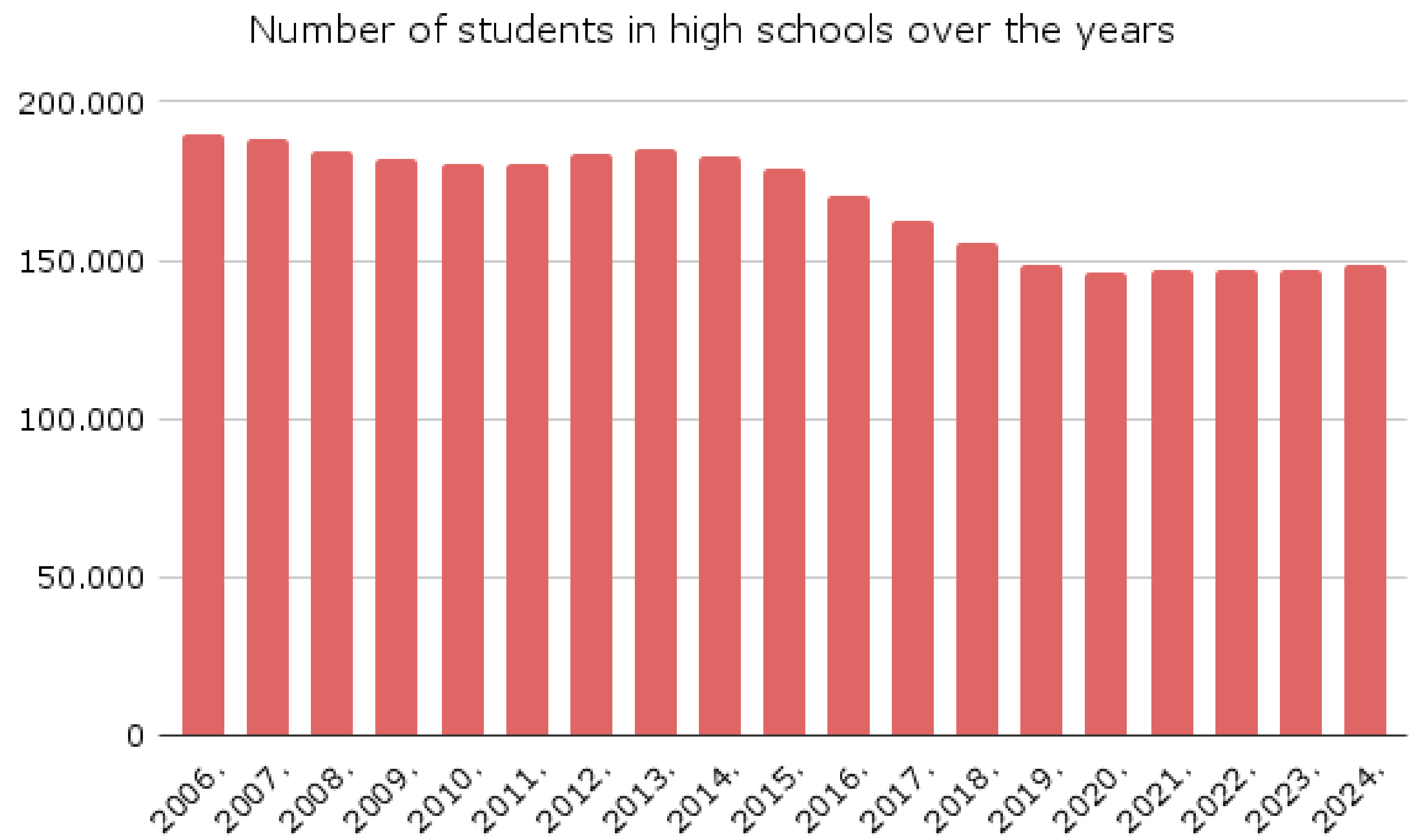
Visualization of Real Data


Number of primary schools over the years


Number of students in elementary schools over the years

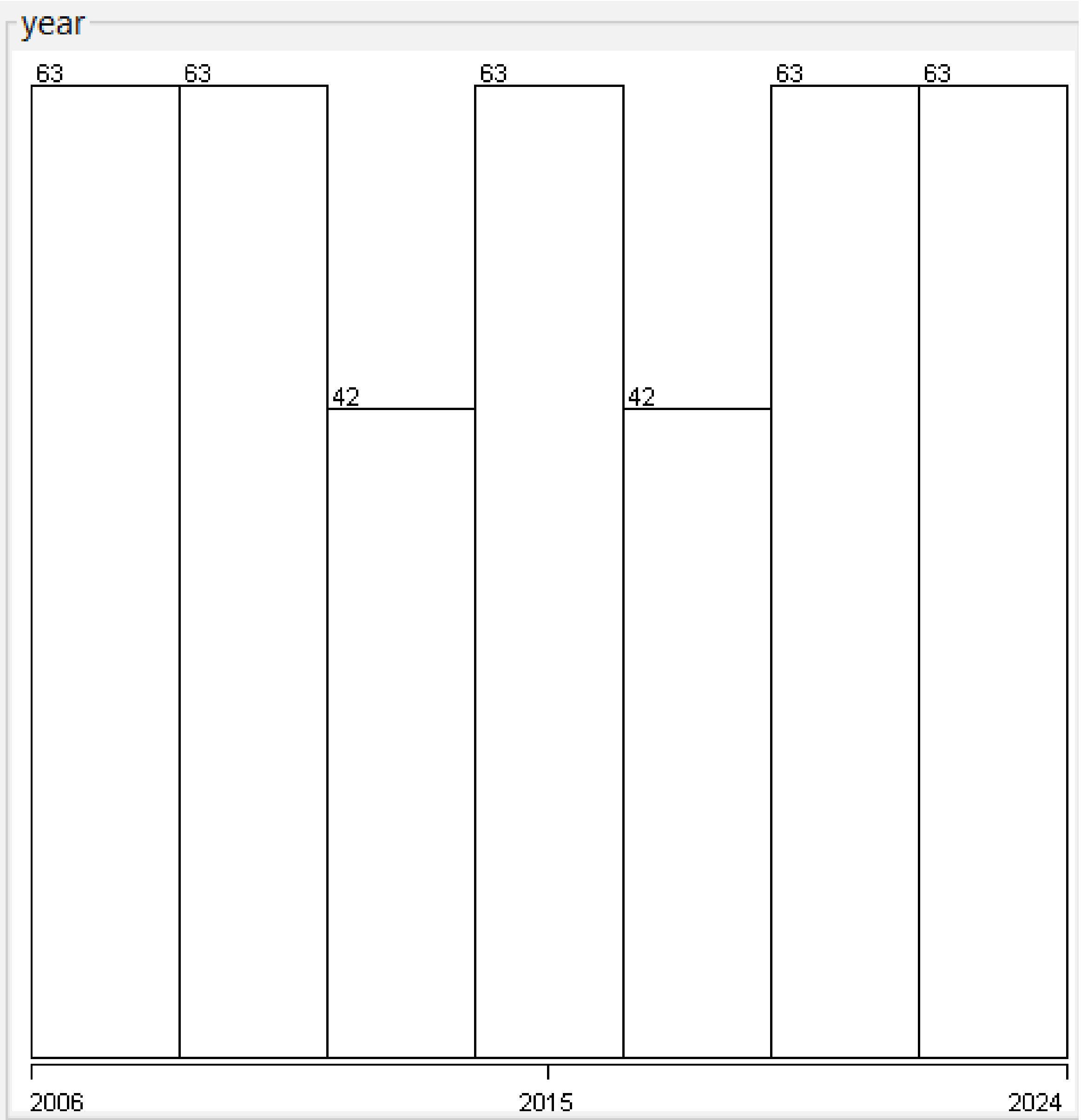
Visualization of Real Data


Number of high schools over the years


Number of students in high schools over the years

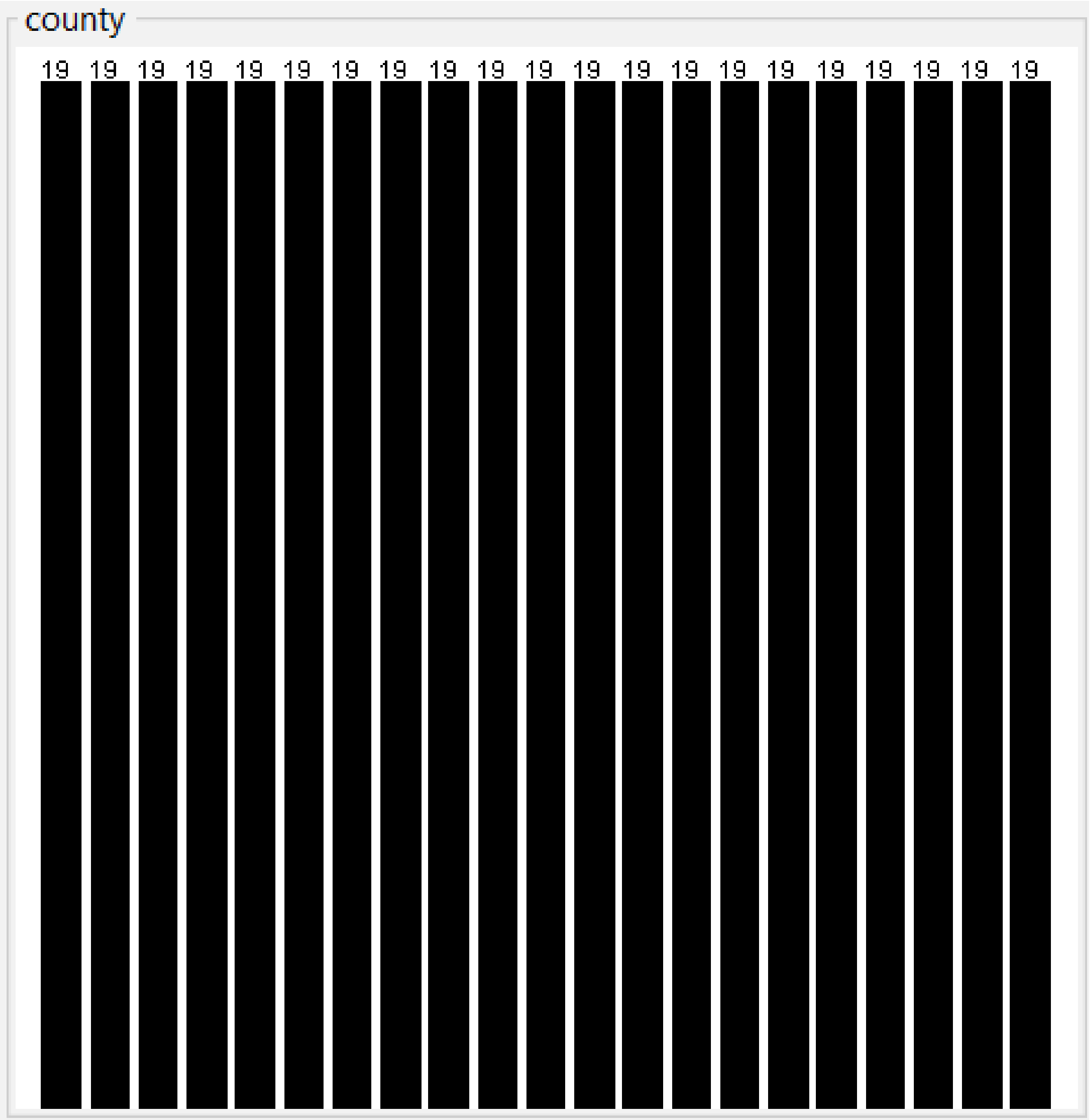
Year

- Data evenly distributed across five time periods
- Slight drop in two specific years
- Supports consistent time-series analysis



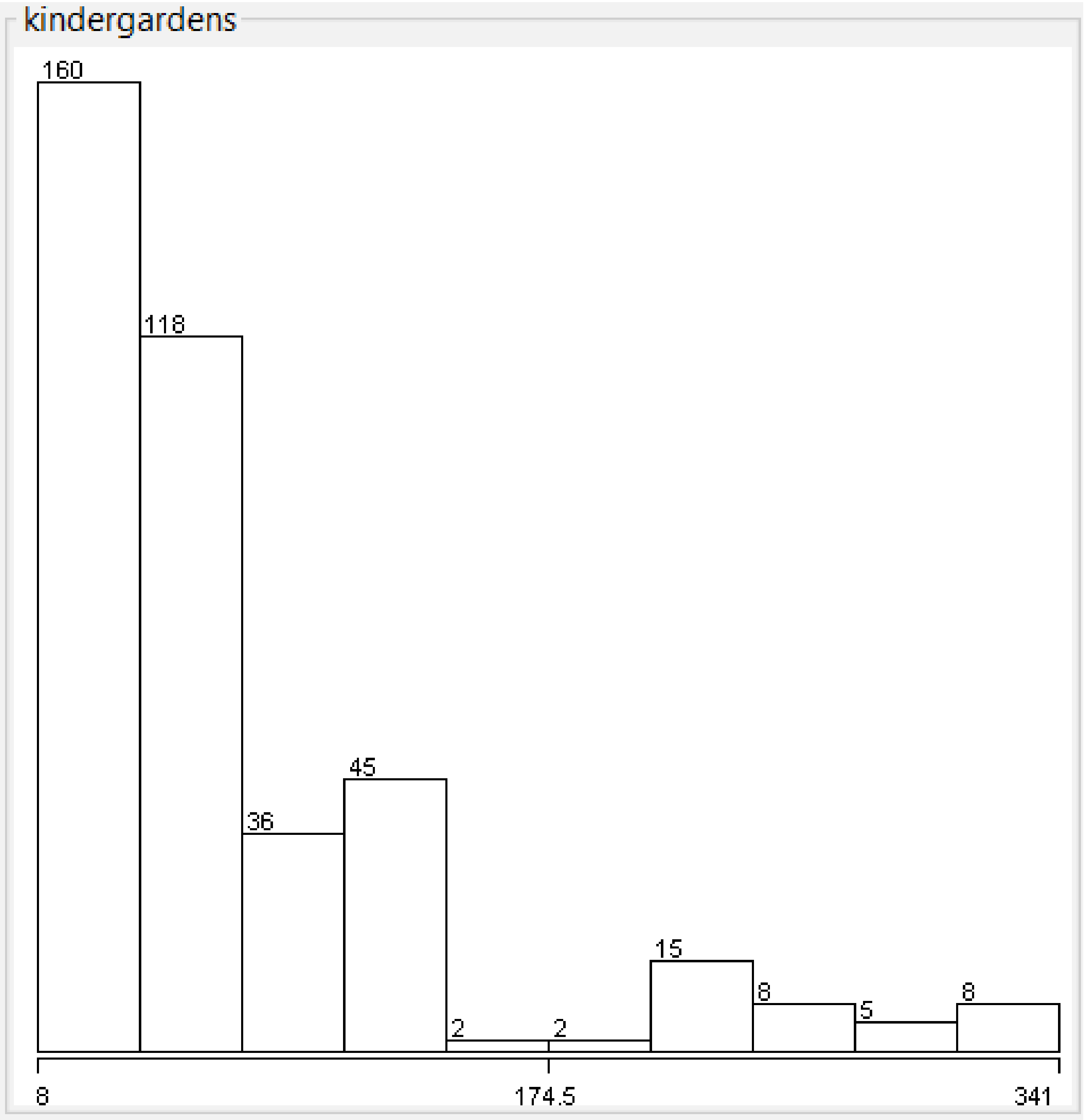
County

- 21 counties, each with exactly 19 records
- Ensures balanced geographic coverage



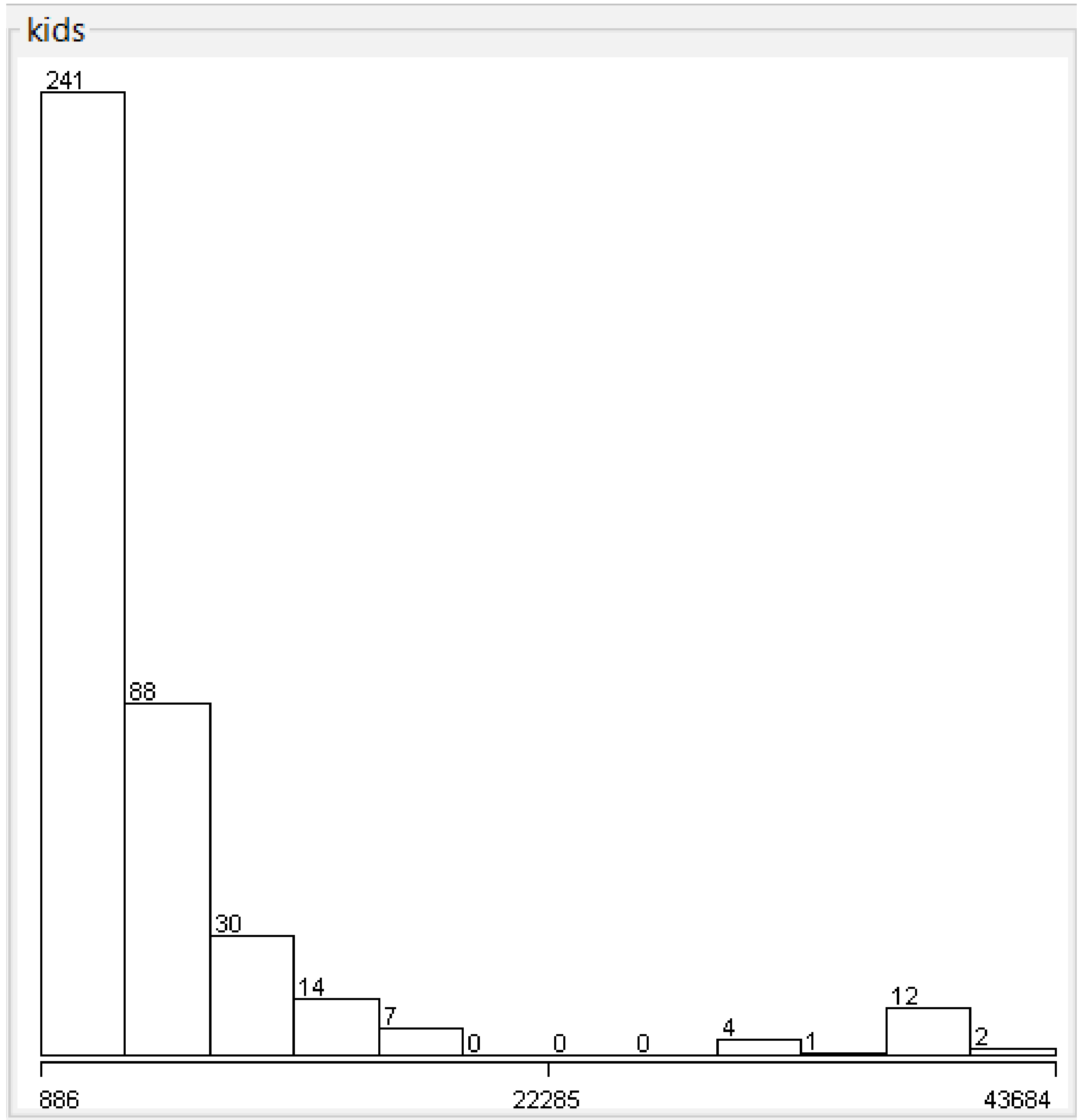
Kindergartens

- Majority of counties have low counts
- Right-skewed distribution
- Few counties with very high values (urban areas)



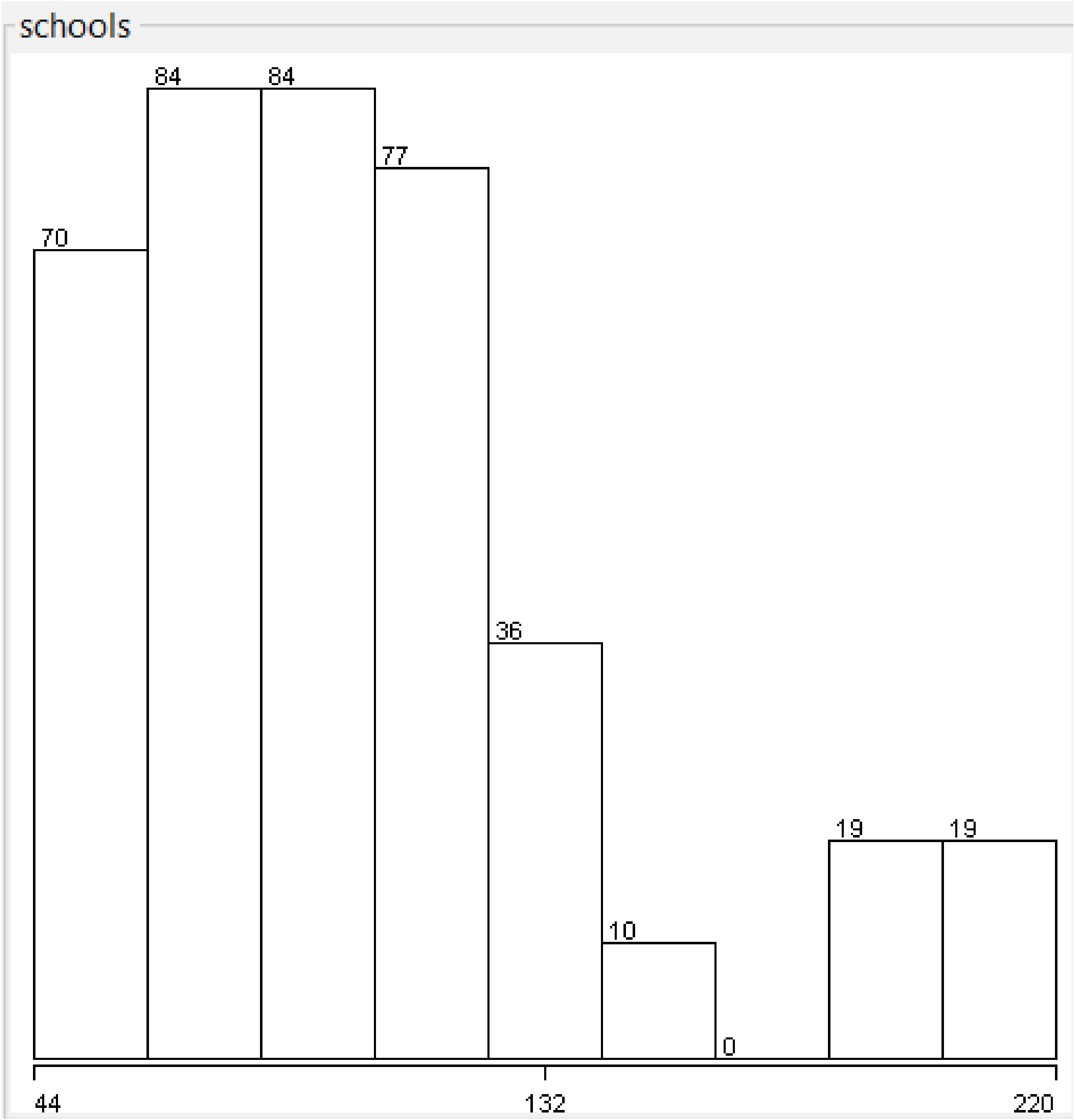
Kids

- Similar distribution to kindergartens
- Most counties under 22,000 children
- A few outliers above 40,000



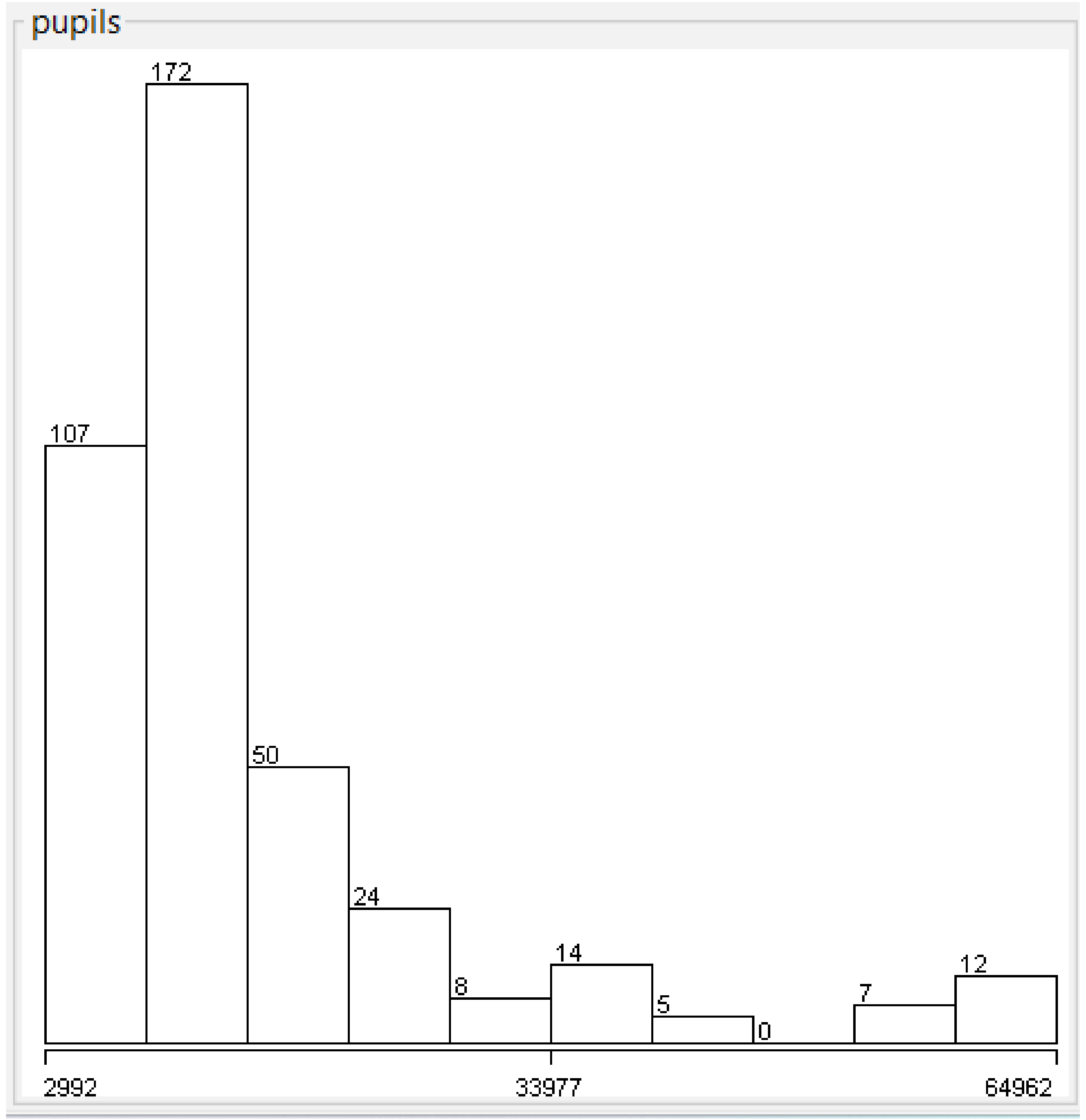
Schools

- Most counties have between 44 and 132 schools
- Some counties exceed 200
- Indicates regional differences in infrastructure



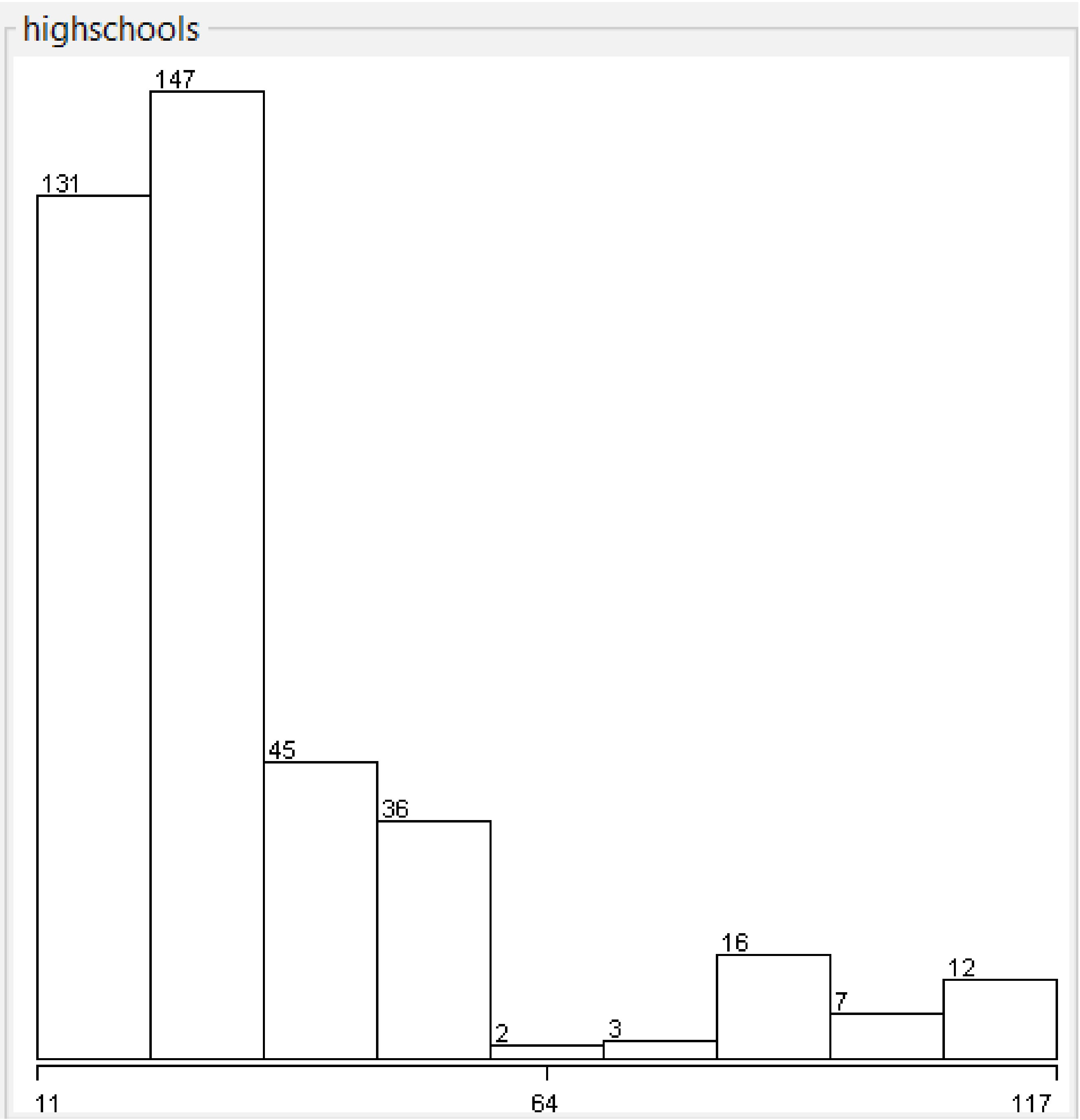
Pupils

- Most counties have under 34,000 pupils
- Strong right-skewed distribution
- Outliers reaching up to 65,000



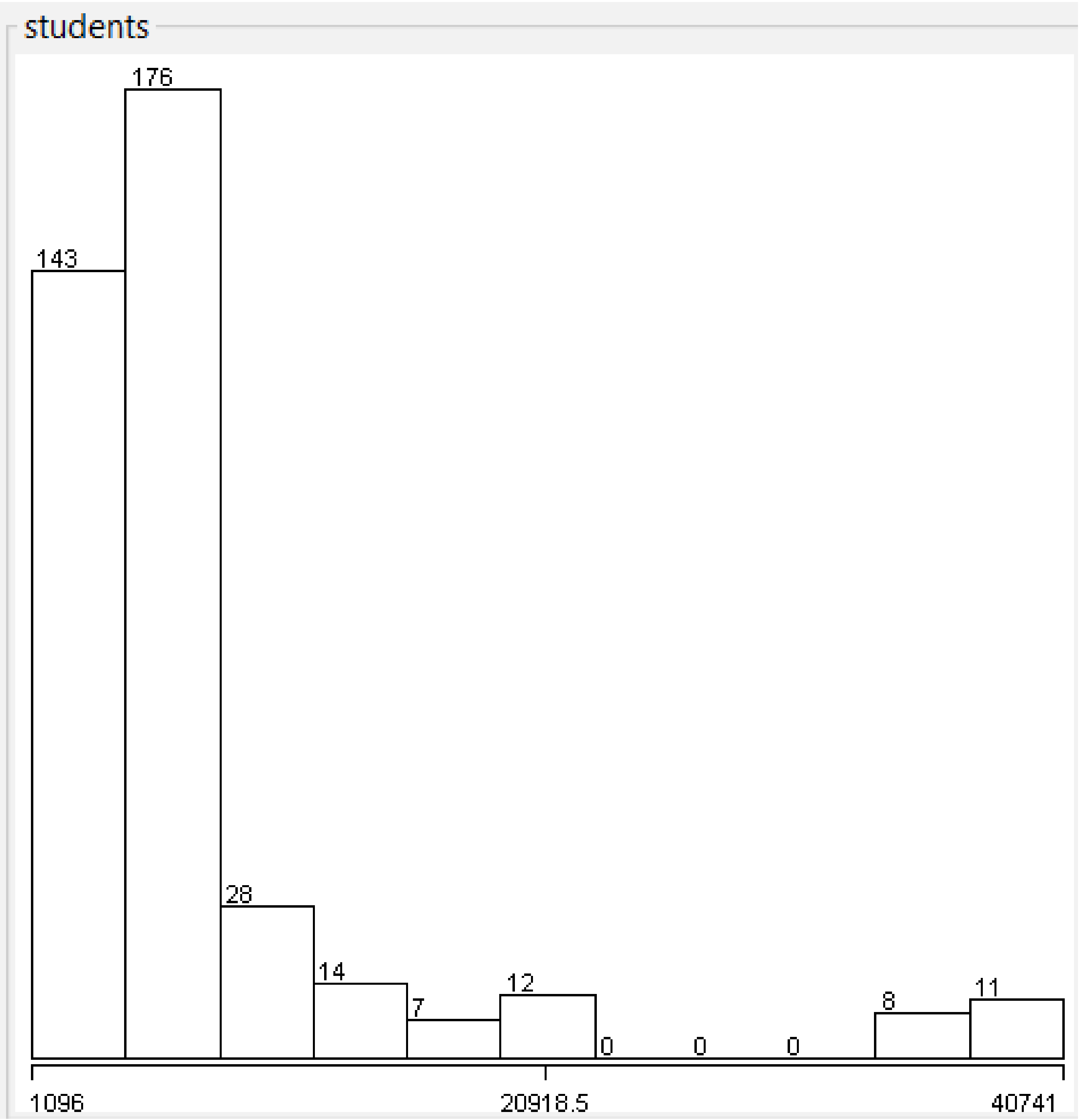
Highschools

- Majority have fewer than 64 high schools
- Only a few counties exceed 100
- Less variation than primary schools



Students

- Most counties have fewer than 20,000 students
- Outliers reach up to 40,000
- Highlights educational demand differences



Data Processing Procedure

- Converted raw .xlsx data into .arff format for use in WEKA
- Removed unnecessary columns (e.g., administrative codes, non-numerical data)
- Converted the form of school year (2005./2006. to 2006.)
- Handled missing or inconsistent values
- Aggregated data by year for national-level prediction
- Encoded categorical attributes if used (e.g., County)
- Normalized numerical values (optional for some algorithms)

Algorithms That WE Used

Linear Regression:

- A simple algorithm that fits a straight line to the data to predict future values.

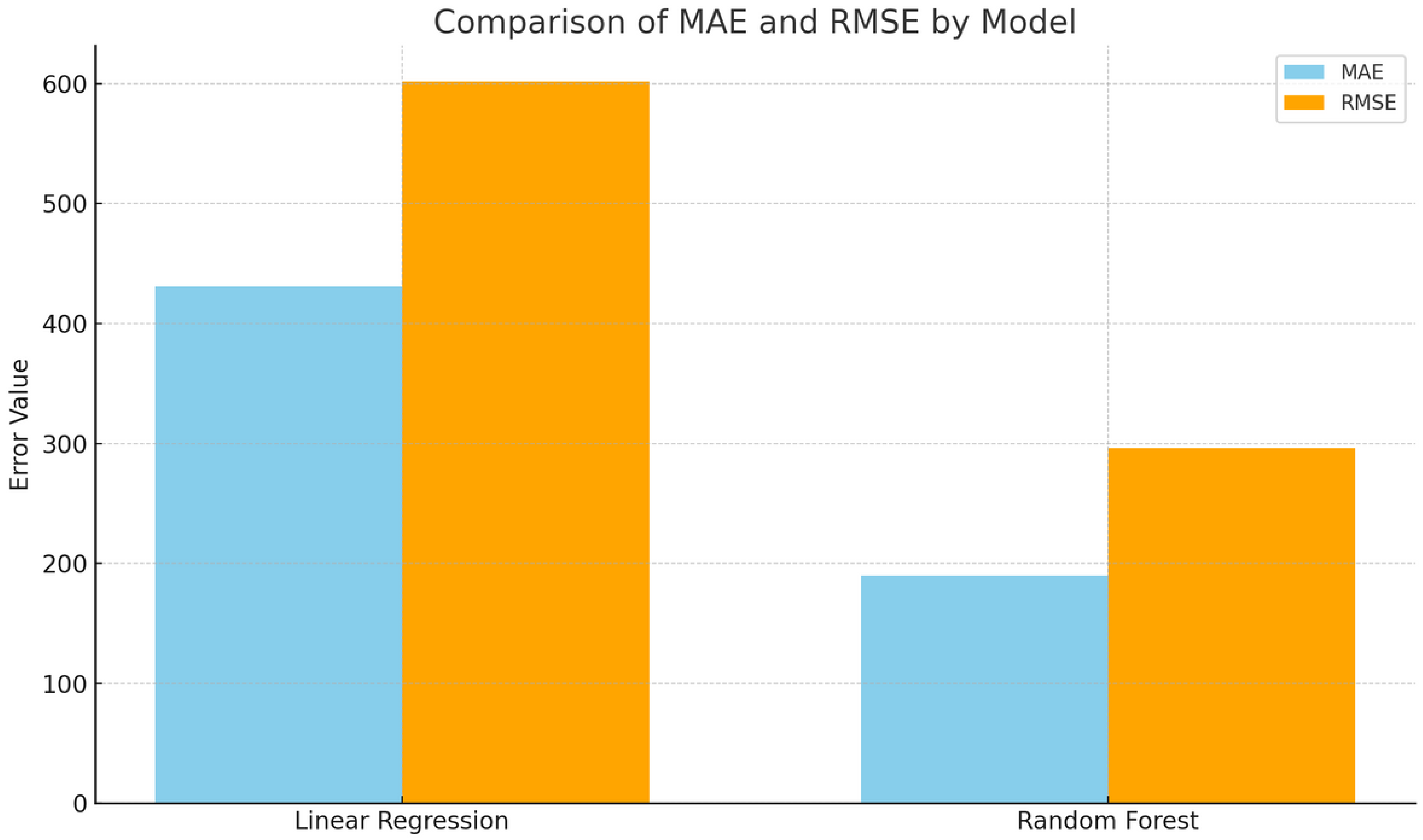
Random Forest:

- An advanced ensemble method that builds many decision trees and averages their results for better accuracy.

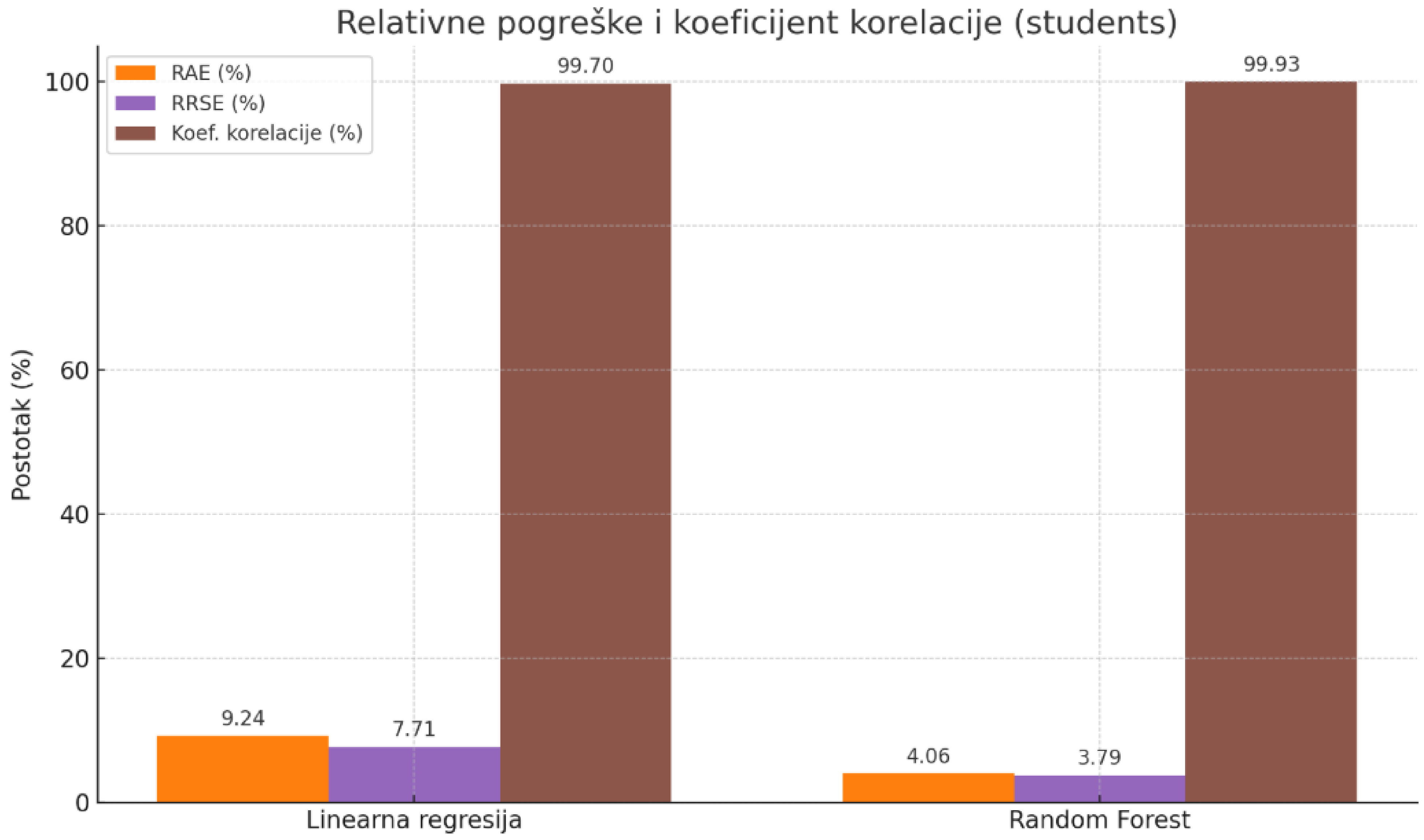
Model Evaluation

Model	Correlation	MAE	RMSE
Linear Regression	0.997	431	602
Random Forest	0.999	190	296

Error Comparison

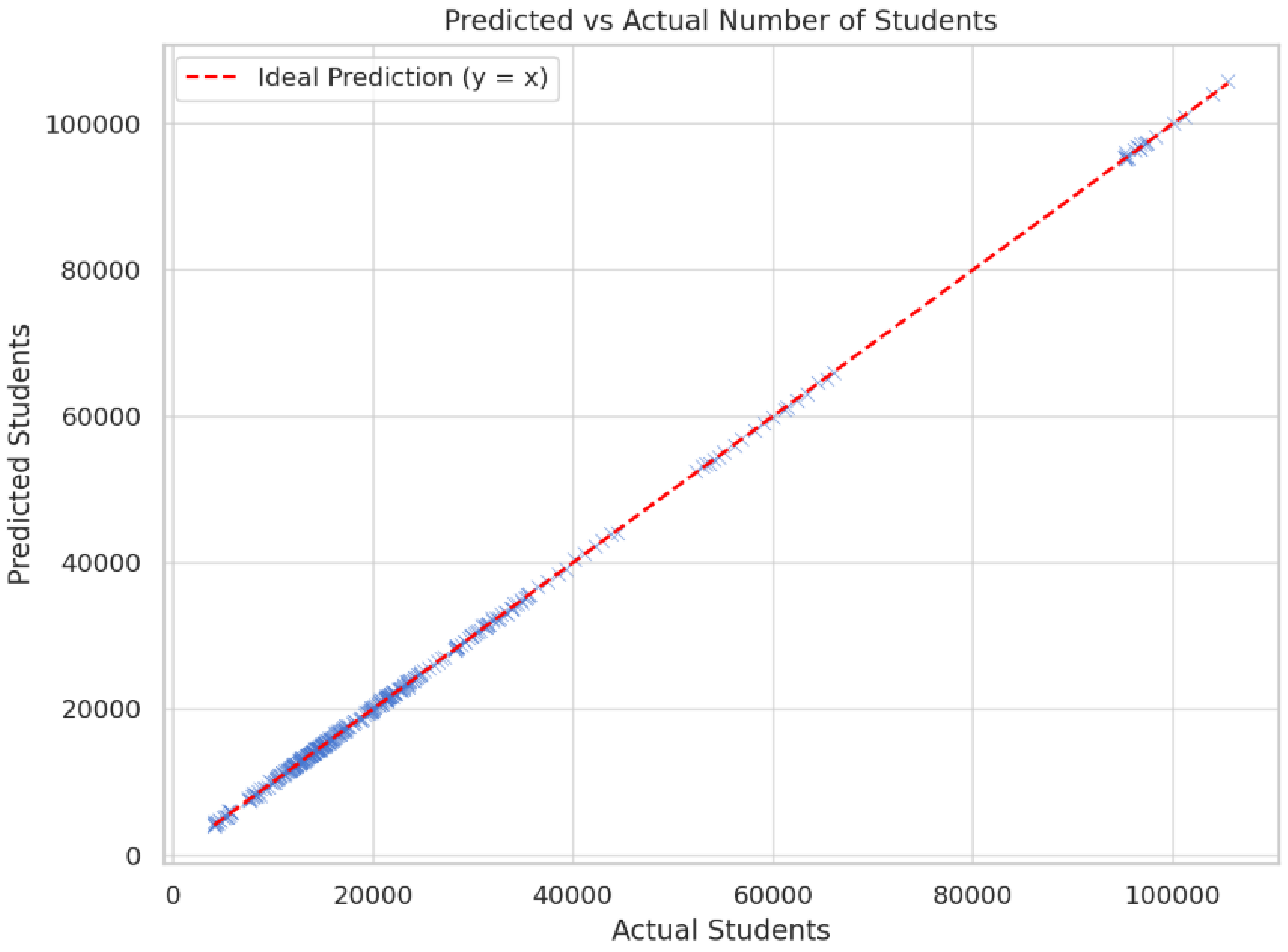


relative errors and correlation coefficient



Scatter Plot: Actual vs. Predicted

This plot shows how closely the predicted student numbers align with the actual student counts. The red dashed line represents the ideal prediction line (perfect match: Predicted = Actual). Most points lie very close to this line, confirming high accuracy of the Random Forest model.

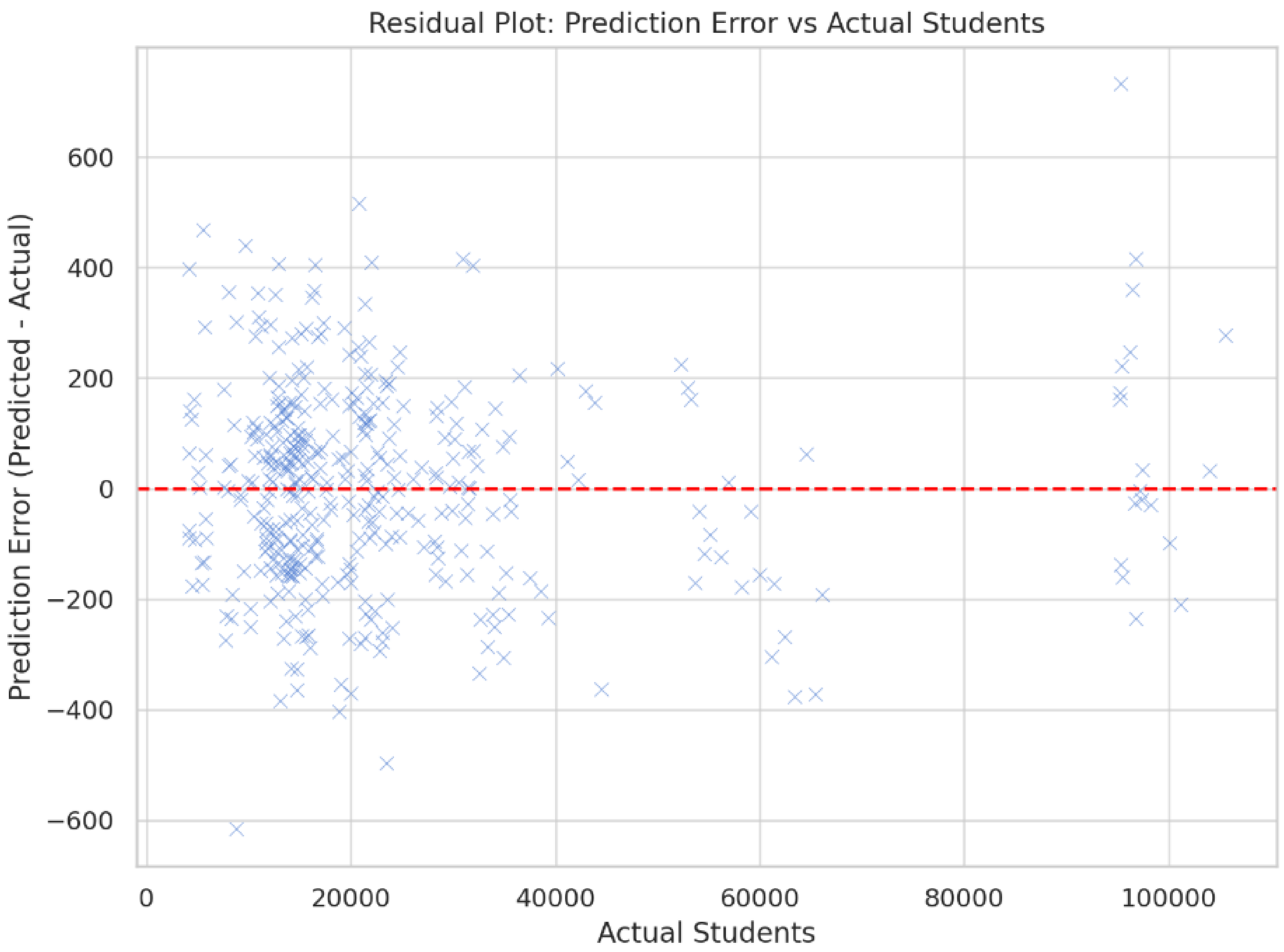


residual plot

This plot shows the difference between predicted and actual student counts. Points are centered around 0, indicating that predictions are mostly unbiased.

Most errors lie within a ±600 range, consistent with the Random Forest model's low mean absolute error (MAE ≈ 190).

A few outliers are present at higher actual student counts, which is expected.

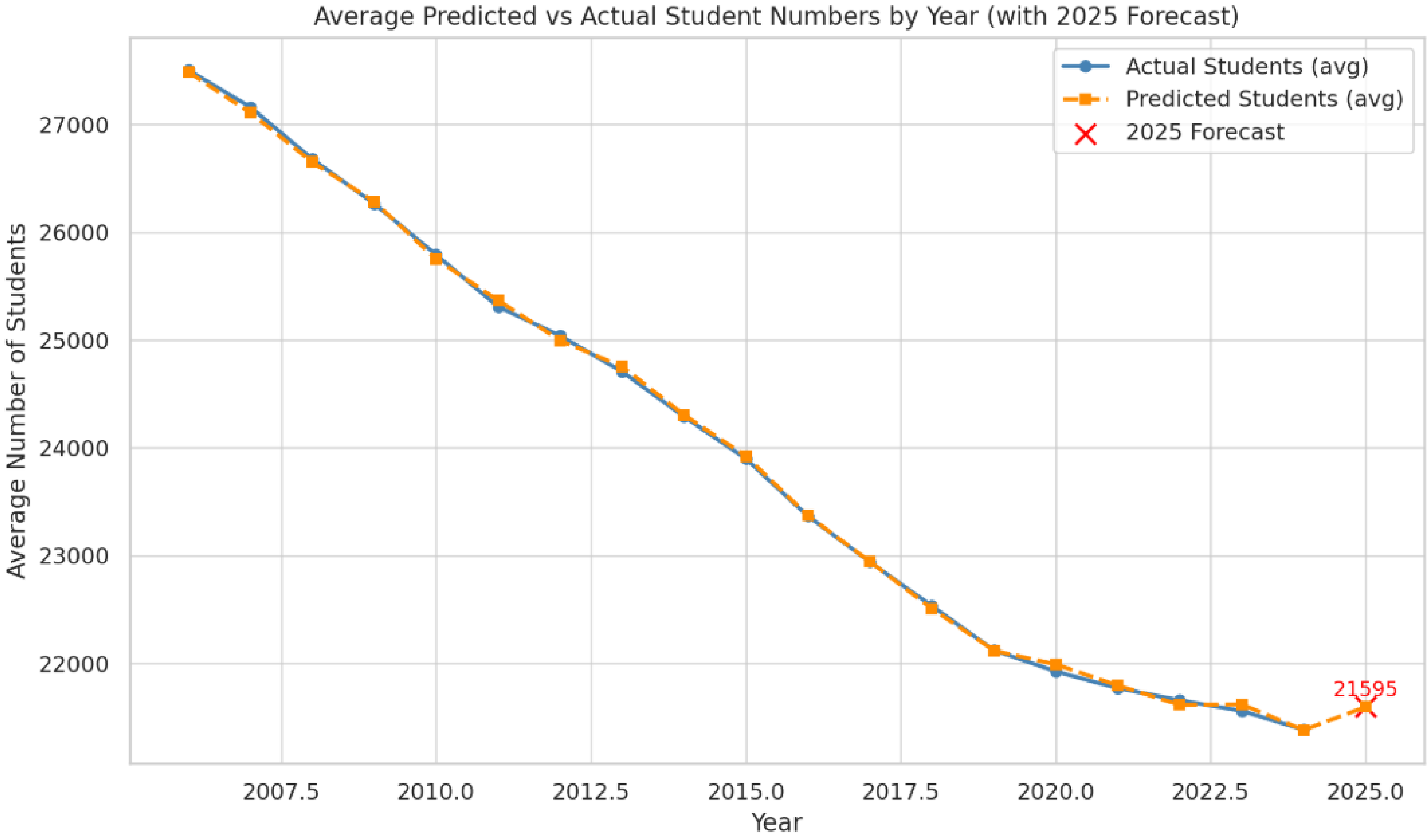


Prediction for 2025.

Shows model-predicted vs actual average student numbers over time.

The red point marks the prediction for 2025 based on recent trends.

2025 value was estimated using a projected 1% annual increase.



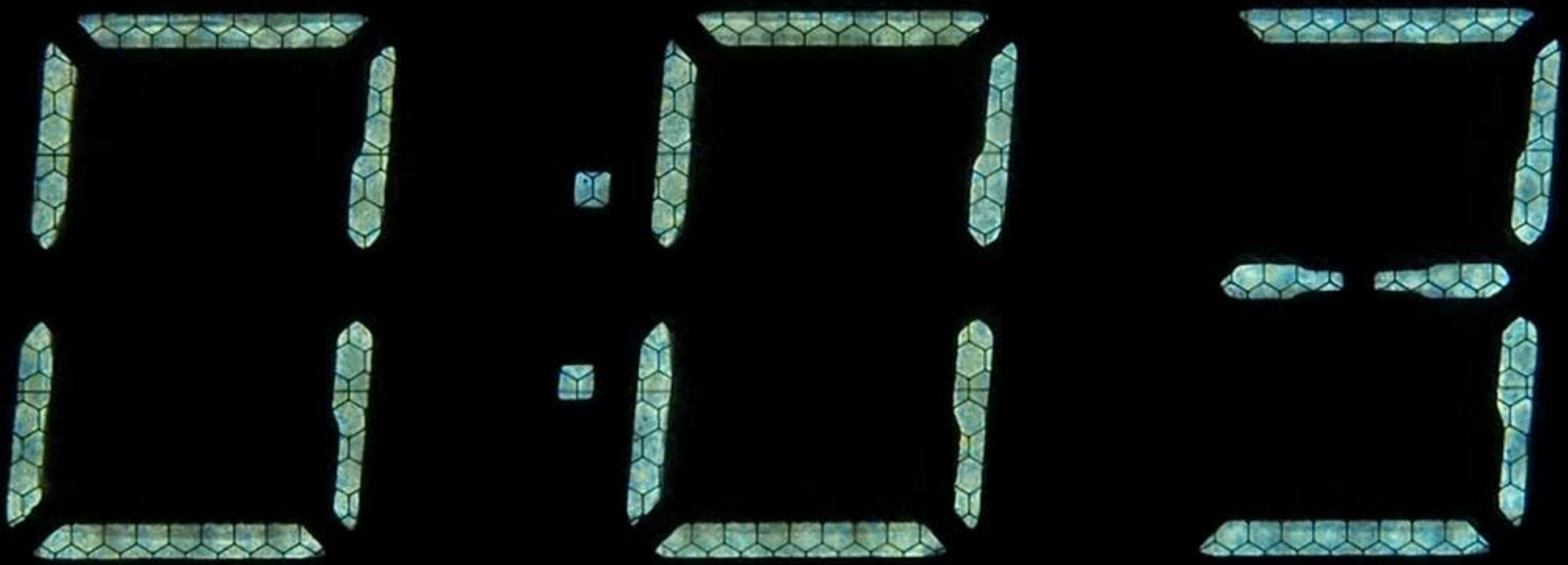
Conclusions

- Random Forest was more accurate than Linear Regression
- WEKA is easy to use for data analysis and prediction
- Predictive models can support education planning
- Visualization helps communicate trends clearly

Application in real life

- Forecast the number of classes, teachers, and classrooms
- Optimize school infrastructure and resource allocation
- Help ministries make data-driven decisions
- Useful in long-term educational strategy and budgeting

One tree may mislead, but a forest, grown from many honest cuts, begins to whisper certainty.


0:03

BIDUSA

Big Data Unites Sciences and Arts

Cofinanciado por
la Unión Europea

Co-funded by the
Erasmus+ Programme
of the European Union