Project Description and Summary

This project aims to leverage keystroke log data to predict overall writing quality. By analyzing process features such as typing speed, revision patterns, and text structure changes, we seek to uncover relationships between writing behaviors and writing performance. Unlike traditional assessment tools that focus solely on the final written product, our approach emphasizes the writing process, potentially enhancing learner autonomy and metacognitive awareness in writing.

Our methodology involved applying a variety of machine learning models to the keystroke log data, including Elastic Net Regression, K-Nearest Neighbors Regression, Random Forest Regression, and Support Vector Machines Classification. Unique predictors like *"avg_cursor_movement"*, *"avg_text_change_length"*, and *"final_word_count"* were pivotal in our analysis. We also utilized unsupervised learning techniques like UMAP and K-Means Clustering, followed by Principal Component Analysis for dimensionality reduction, to better understand the natural clustering of the data.

Apart from the standard models, we employed Elastic Net Regression for its dual ability to handle multicollinearity and perform variable selection, crucial for our dataset with highly correlated variables. The application of Principal Component Analysis for dimensionality reduction before applying K-Means Clustering is another specialized method. These approaches differ from standard methods by offering a more nuanced understanding of the data structure, crucial for handling our dataset's high dimensionality and intricate variable interactions.

Our findings revealed that Random Forest Regression yielded the most accurate predictions (lowest RMSE), highlighting the importance of variables like *"final_word_count*", "*total_events*", and "*max_cursor_movement*". However, clustering models showed limited efficacy due to the data's complexity, underscoring the challenges in modeling such intricate datasets. This project demonstrates the potential of using process features in writing assessment, paving the way for more nuanced and formative writing evaluation tools.