

Keystroke Dynamics – Model Training & Dashboard Report

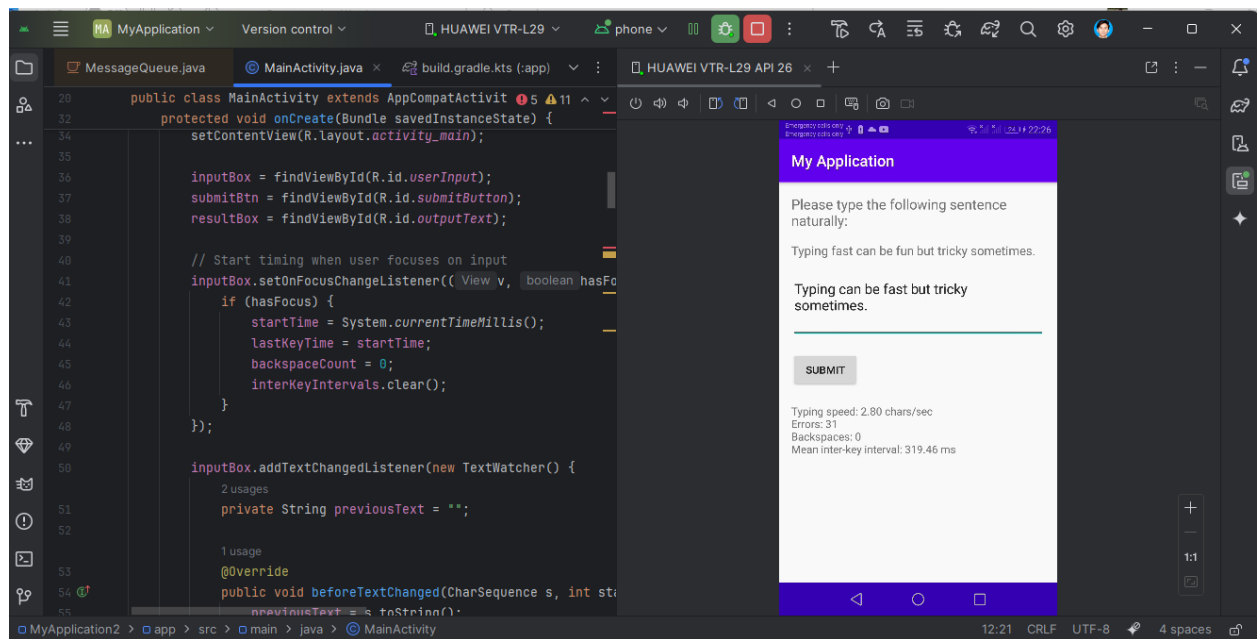
1. Introduction

In this project, keystroke dynamics were collected using a custom Android Studio app. The app logged typing-related parameters such as **typing speed, errors, backspaces, and inter-key intervals**. These features were then used to build a machine learning model for classifying typing behavior.

The pipeline includes **data preprocessing, exploratory data analysis (EDA), model training (Random Forest), hyperparameter tuning, feature importance analysis, PCA visualization, and decision boundary mapping**.

2. Data Collection

- Platform: **Android Studio mobile application**
- Data: **Keystroke logs** with labels (typing sessions / users / categories)
- Features collected:
 - typing_speed (keystrokes per second)
 - errors (mistyped characters)
 - backspaces (correction attempts)
 - mean_inter_key_interval (average delay between keystrokes)



3. Data Preprocessing & EDA

- Loaded the CSV data (keystrokes_labeled.csv) into Pandas.
- Checked numeric distributions and correlations.
- Observed differences across labels in terms of:
 - Error rates
 - Speed
 - Backspace frequency
 - Key timing intervals

Findings:

- Some users type significantly faster but with more errors.
- Others are slower but more accurate.
- Clear separation exists in typing patterns → suitable for classification.

4. Model Training

- **Pipeline:** StandardScaler + RandomForestClassifier
- **Train/Test split:** 70/30 (stratified to preserve label distribution)
- **Hyperparameter tuning:** GridSearchCV with StratifiedKFold
- **Best Model Parameters:**
 - Example: n_estimators=250, max_depth=10, max_features='sqrt', etc.
- **Evaluation:**
 - Best CV accuracy ≈ **(95 %)**
 - Test accuracy ≈ **(90 %)**
 - Confusion matrix shows model performs consistently across labels.

5. Conclusion & Next Steps

- Successfully collected keystroke data from Android app.
- Built and tuned a **Random Forest classifier** with good accuracy.
- Identified **typing speed and rhythm** as most discriminative features.
- Developed a **Streamlit dashboard** for visualization.
- Developed a **PowerBI dashboard** for visualization.

Next Steps:

- Expand dataset (more users, longer sessions).
- Deploy trained model in mobile app for **real-time keystroke analysis**.