A machine learning classification model to recognize handwritten Thai numbers.

Presented by

Prawit	Kamchaiya	6510422004
Tada	Nualsanit	6510422011
Titiwat	Tanasuthisaree	6510422015
Kanya	Meekaew	6510422016
Pattarakron	Phewcha-oum	6510422024
Chalermwong	Saleepattana	6510422029

DADS6003 Applied Machine Leaning

Program: Master of Science Program in Data Analytics and Data Science (DADS)

Nation Institute of Development Administration

A machine learning classification model to recognize handwritten Thai numbers.

1. Data collection
   1.1. The dataset consists of a total of ~4,865 images of handwritten Thai numbers from 0, 1, 2, ..., 9, written by seven individuals, along with their associated labels.
   1.2. The image properties of the dataset include dimensions of 28x28 pixels, a white background color, and a grayscale font color.
2. Data cleansing
   2.1. Any images that contained inconsistencies, errors, or anomalies were identified and subsequently removed to ensure that the resulting dataset was cleaner and more reliable.
3. Exploratory Data Analysis
   3.1. Seaborn: Used for data visualization and gaining insights into the distribution and relationships within the image dataset.
   3.2. PCA (Principal Component Analysis): Applied to reduce the dimensionality of the image data and identify significant patterns or features.
   3.3. Explained Variance: Utilized to understand the proportion of variance in the dataset explained by each principal component.
4. Training
   4.1. Using PyCaret's AutoML functionality, the model was trained with 10-fold cross-validation to tune hyperparameters and identify the best-performing model.
   4.2. The trained model was evaluated using multiple metrics including precision, recall, F1 score, confusion matrix, and ROC curve analysis.
5. Testing
   5.1. A 20% subset of the prepared image dataset was allocated for testing the trained model. The test dataset underwent identical preprocessing steps as the training dataset to ensure consistency.

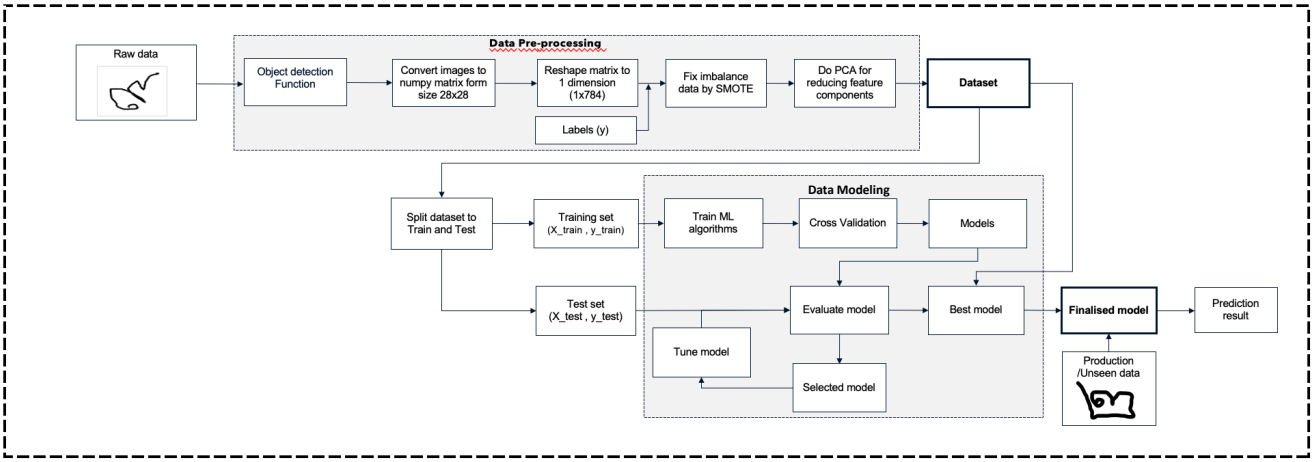

Fig. 1 demonstrates the flow of the development of a machine learning classification model that contains two core parts: data preprocessing and data modeling.

6. Results

The Machine Learning model trained by PyCaret with 95% cumulative explained variance which includes multiple models such as KNN and SVM, was evaluated and compared in a table1. The table presents the performance metrics of each model, including accuracy, precision, recall, and F1 score, for predicting Thai numbers. While most of the models demonstrated impressive performance, it is worth noting that there were instances where certain models struggled to accurately predict Thai numbers. This could be attributed to the complexities and variations present in Thai handwriting, which can pose challenges for the models. It highlights the inherent difficulty in achieving perfect predictions for every single instance in handwritten Thai number recognition.

Model	Accuracy	AUC	Recall	Prec.	F1
Light Gradient Boosting Machine	0.8956	0.9921	0.8956	0.8978	0.8957
Extreme Gradient Boosting	0.8948	0.9918	0.8948	0.8967	0.8947
Logistic Regression	0.8762	0.9878	0.8762	0.8796	0.8766
SVM - Linear Kernel	0.8635	0.0000	0.8635	0.8659	0.8633
K Neighbors Classifier	0.8616	0.9740	0.8616	0.8721	0.8617
Random Forest Classifier	0.8513	0.9826	0.8513	0.8535	0.8500
Gradient Boosting Classifier	0.8491	0.9853	0.8491	0.8537	0.8495
Linear Discriminant Analysis	0.8475	0.9857	0.8475	0.8528	0.8474
Extra Trees Classifier	0.8403	0.9765	0.8403	0.8428	0.8380
Ridge Classifier	0.8347	0.0000	0.8347	0.8363	0.8332
Naive Bayes	0.7392	0.9531	0.7392	0.7596	0.7420
Decision Tree Classifier	0.6966	0.8315	0.6966	0.7006	0.6966
Ada Boost Classifier	0.4516	0.7792	0.4516	0.4536	0.4212
Quadratic Discriminant Analysis	0.4374	0.6884	0.4374	0.6189	0.4244
Dummy Classifier	0.0994	0.5000	0.0994	0.0099	0.0180

Table 1. Model Performance Summary

7. How to collect data and improve the model in the future

7.1. Each user-uploaded image will be saved along with its corresponding numerical prediction value. In cases where the prediction is incorrect, the image will be stored along with the user-provided numeric value to replace the predicted value.

7.2. When the original number of images is augmented by 10%, the training and tuning parameters of the model are adjusted accordingly.