KNIME CLASSIFICATION PROJECT

SVM vs MLP: Data Science Workflow & Evaluation

A. DATA EXPLORATION & VISUALIZATION

- - Used boxplots, histograms, and statistics nodes
- - Dataset: 210 samples, 3 classes
- - Detected zero-variance features and class imbalance
- - Partial linear separability observed via PCA

B. FEATURE IMPORTANCE

- - Correlation Matrix and Information Gain used
- - High-correlation features selected
- - Redundant and low-entropy features removed

C. LINEAR SEPARABILITY ANALYSIS

- - Used scatter plots and PCA
- - Classes partially separable
- - Justified use of non-linear classifiers like MLP

D. KNIME WORKFLOW & NODE CONFIGURATION

- Used CSV Reader, Missing Value Handler, Normalizer

- Learners: SVM, MLP

- X-Partitioner + Aggregator: Cross-validation

- PMML Writer: Exported models

E. IMPORTANCE OF TRAIN-TEST SPLIT

- - Ensures generalization and prevents overfitting
- - Reflects real-world unseen data prediction
- - Supported by Goodfellow et al. (2016)

F. LEARNER VS PREDICTOR IN KNIME

- Learner: Trains model on labeled data

- Predictor: Applies model to test data

- Core nodes for SVM and MLP workflows

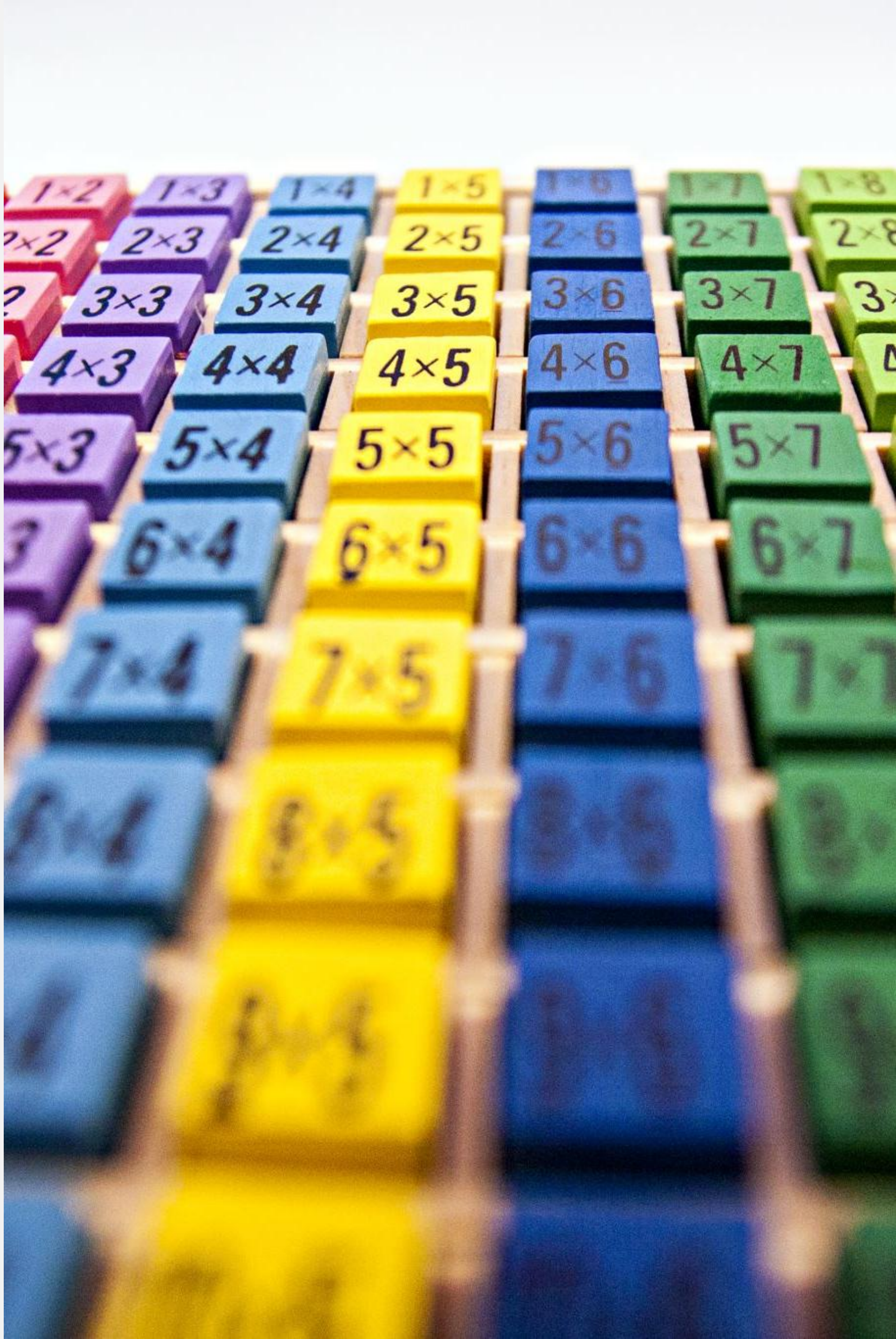
G. HOW SVM & MLP WORK

- SVM:
- - Finds optimal hyperplane
- - Supports linear & non-linear via kernels

- MLP:
- - Neural network with hidden layers
- - Learns complex functions
- - Uses backpropagation

H. RESULTS & ANALYSIS

- - Accuracy: 100% for both models
- - Precision/Recall/F1: 1.0 across all classes
- - Confusion Matrix: Perfect prediction
- - Caution: small dataset size may overstate results



I. HYPERPARAMETER OPTIMIZATION

- Used parameter optimization loop
- SVM: kernel type, C
- MLP: learning rate, hidden layers
- Improved convergence and generalization

J. K-FOLD CROSS VALIDATION

- - Every sample used in training & testing
- - Reduces variance in metrics
- - Implemented using X-Partitioner & Aggregator
- - Reference: Kohavi (1995)

CONCLUSION & TAKEAWAYS

- - Complete supervised learning pipeline built
- - SVM & MLP tuned and validated
- - Results excellent, but require larger data
- - Future: ensemble models, real-time deployment