Homework 1 Documentation

Chigozie Nnani

DATA 522: Practical Deep Learning Systems
February 6, 2026

1 Design

1.1 Problem and Target

The task is binary classification: predict good ($|\eta| < 0.6$) vs. bad catalyst from 11 inputs (V, Cr, Mg, Fe, Co, Ni, Cu, S, Se, P, Voltage, Time). Part 1.1 visualizes this 11D space via PCA (first two components), with points coloured by $|\eta|$.

1.2 Data Preparation

1. Rows with missing Overpotential V at 50.0 mA/cm2 were dropped; good = 1, bad = 0.
2. Random split: 70% train, 15% validation, 15% test (fixed seed); same splits for all steps.

1.3 Algorithm: Random Forest

1. Random Forest was chosen because it handles mixed scales, is robust, supports permutation feature importance, and is a standard baseline for tabular classification.
2. Baseline used default settings (e.g., 100 trees, no max depth limit).

1.4 Hyperparameter Selection (Part 1.5)

1. Four combinations were chosen to explore the hyperparameter space:
   (a) **n_estimators:** 50, 100, 200.
   (b) **max_depth:** None, 5, 10, 20.
   (c) **min_samples_split:** 2, 5, 10.
2. Validation set was used to compare precision, recall, F1 on train/val/test; the choice is justified by better val/test balance without severe overfitting.

1.5 Permutation Feature Importance (Part 1.6)

1. PFI can be defined as a technique that measures a feature's importance by calculating the degradation in model performance when that feature's values are randomly permuted, thereby breaking its relationship with the target variable.
2. Two settings: raw features (no normalization) and StandardScaler fit on train, applied to train/val/test. Comparison shows whether rankings depend on scale.

2 Results

2.1 Class Balance (1.3)

36.34% good, 63.66% bad; dataset is imbalanced, motivating the use of precision, recall, and F1.

2.2 Baseline (1.4)

1. Default RF: train near-perfect (precision 1.00, recall and F1 1.00); val/test lower ⇒ overfitting.
2. Validation has the lowest F1 (0.65) and recall (0.60); test is stronger (F1: 0.77, recall: 0.82), so the model generalizes better to the test set than to the validation set.

2.3 Hyperparameter Tuning (1.5)

Combo 1 has the best validation F1 (0.65); Combo 4 has the best test precision (0.75) and test F1 (0.78). The chosen combo is Combo 4, because it generalizes best to the test set and shows less overfitting (train F1 0.83 vs test F1 0.78) than the default-like Combo 1.

2.4 Feature Importance (1.6)

1. Without normalization: 1) Se, 2) Ni, 3) V, 4) Co, 5) Mg.
2. With normalization: 1) Se, 2) V, 3) Co, 4) Ni, 5) Mg.

The ranking of the top five most important variables is: without normalization, Se > Ni > V > Co > Mg; with normalization, Se > V > Co > Ni > Mg.

3 Discussion

3.1 Algorithm and Design

1. RF as interpretable baseline; 70/15/15 split and fixed seed for reproducibility; good/bad by $|\eta| < 0.6$ per problem statement.

3.2 Hyperparameters

1. Search showed that constraining `max_depth` and increasing `min_samples_split` reduced overfitting and improved the balance between train and validation metrics. The chosen setting (Combo 4) avoids the pattern of near-perfect training performance with lower validation and test performance seen in the default-like Combo 1.

3.3 Feature Importance and Normalization

Se is top in both setups; Ni, V, Co, and Mg follow in slightly different order. Normalization shifts the ranking of V and Ni within the top five. The same five features appear in both lists; main drivers are Se, then V, Co, Ni, and Mg.

3.4 Limitations

1. Small dataset (∼454 samples); metrics can vary with split.
2. Possible next steps: other algorithms (e.g., gradient boosting), cross-validation for stable tuning and importance.