Portfolio Assessment 1 - Decision Tree Classification Report

Combined Cycle Power Plant (CCPP) Dataset

Student details

Name: Phan Cong Hung

Student number: 104995595

Dataset selected

Dataset: Combined Cycle Power Plant (CCPP), file used: Folds5x2_pp.xlsx

Predictors: AT (Ambient Temperature), V (Exhaust Vacuum), AP (Ambient Pressure), RH (Relative Humidity)

Target (original): PE (Net hourly electrical energy output)

Dataset link/source: CCPP

Reason for choosing the dataset

The Combined Cycle Power Plant dataset was selected because it represents a real-world engineering system with a clear measurable target (PE) and a small set of meaningful environmental predictors. Its sample size is sufficient for train/test evaluation, and its numeric structure supports EDA-driven feature selection, feature engineering, and decision-tree classification experiments.

Data preparation and cleaning

Data understanding

The dataset contains 9568 rows and 5 numeric variables (AT, V, AP, RH, PE). All variables are continuous numerical measurements.

Cleaning steps

Duplicate removal: 41 duplicate rows removed.

Missing values: no missing values detected across all columns.

Outlier removal: outliers were identified using the IQR (1.5xIQR) rule across all variables, removing 104 rows.

Final cleaned dataset size: 9423 rows x 5 columns.

Exploratory Data Analysis (EDA)

Descriptive statistics (cleaned data)

PE summary (cleaned): mean=454.36, std=17.1, min=420.26, max=495.76.

Correlation with target (PE)

- AT vs PE: r=-0.95
- V vs PE: r=-0.87
- AP vs PE: r=0.52
- RH vs PE: r=0.39

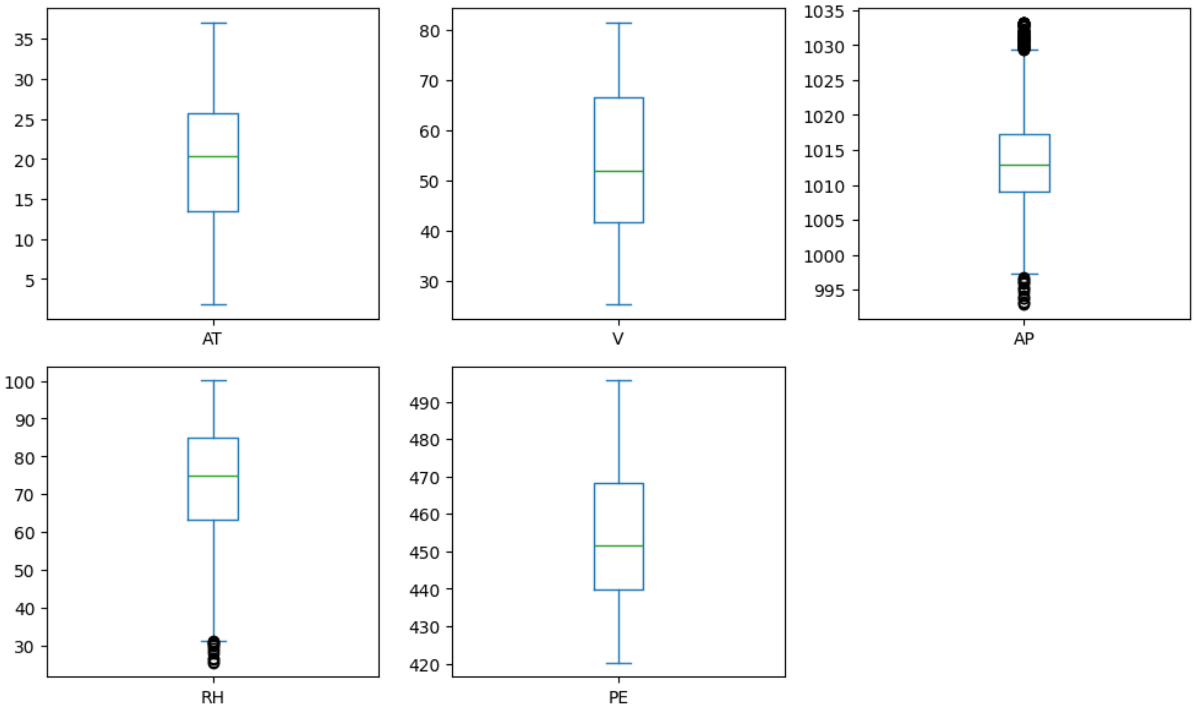

Figure 1: Boxplots of AT, V, AP, RH, and PE (before outlier removal).

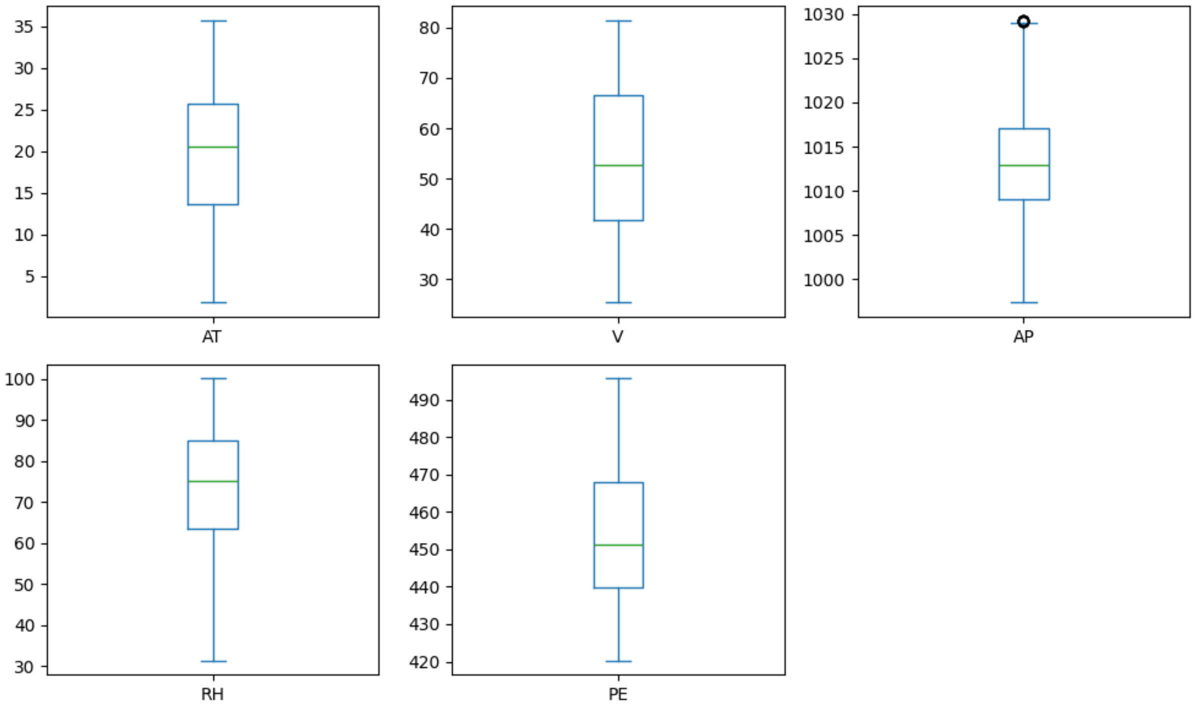

Figure 2: Boxplots of AT, V, AP, RH, and PE (after outlier removal).

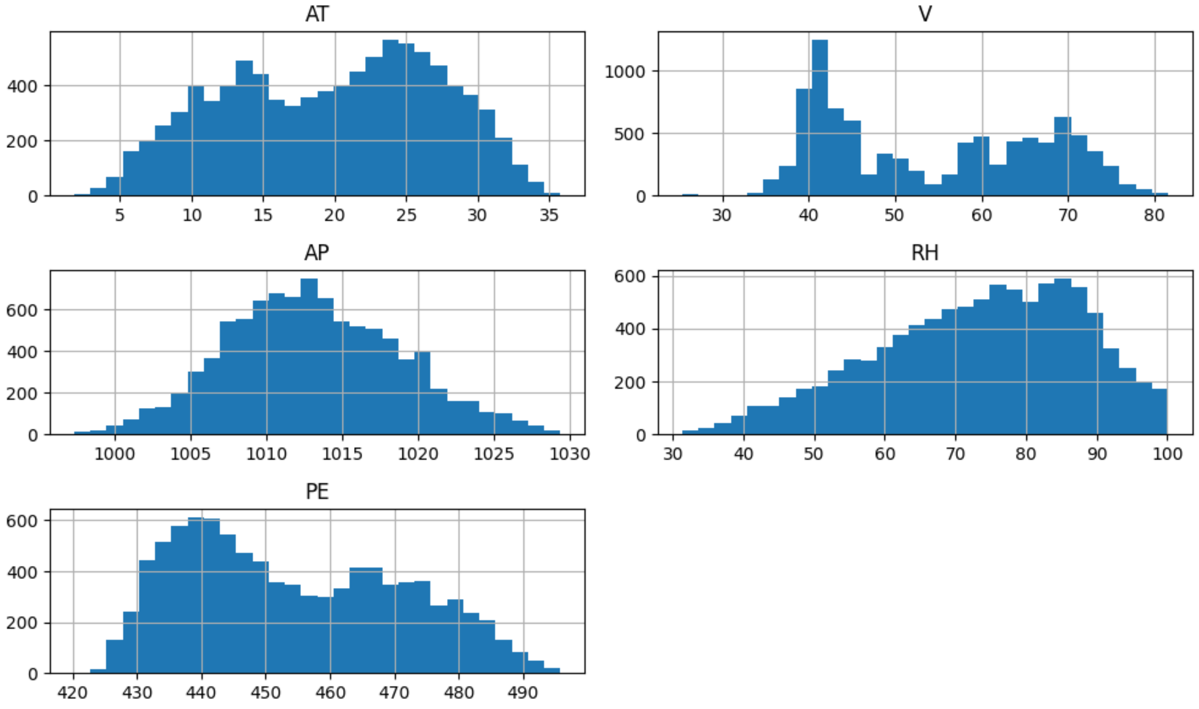

Figure 3: Histograms of AT, V, AP, RH, and PE (cleaned dataset).

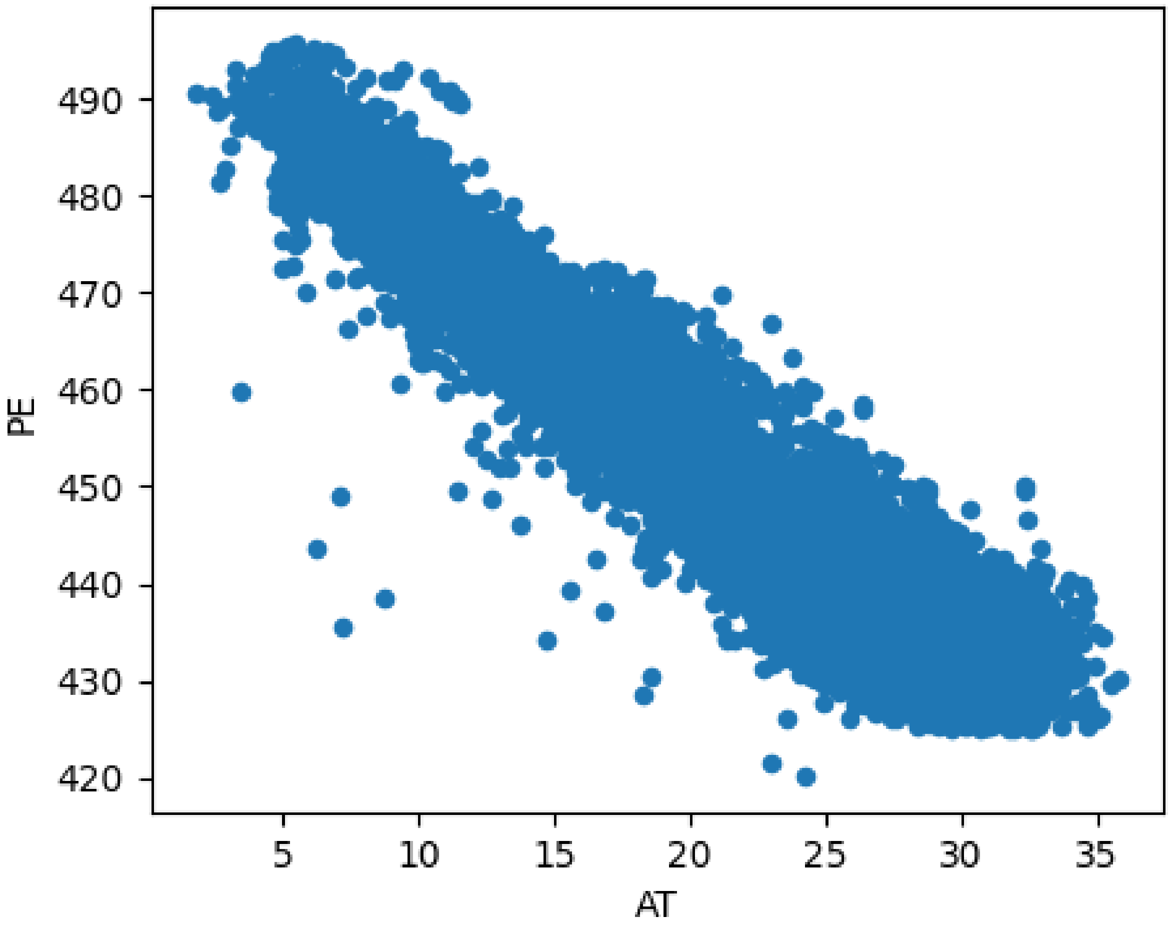

Figure 4: Scatter plot of AT vs PE (cleaned dataset).

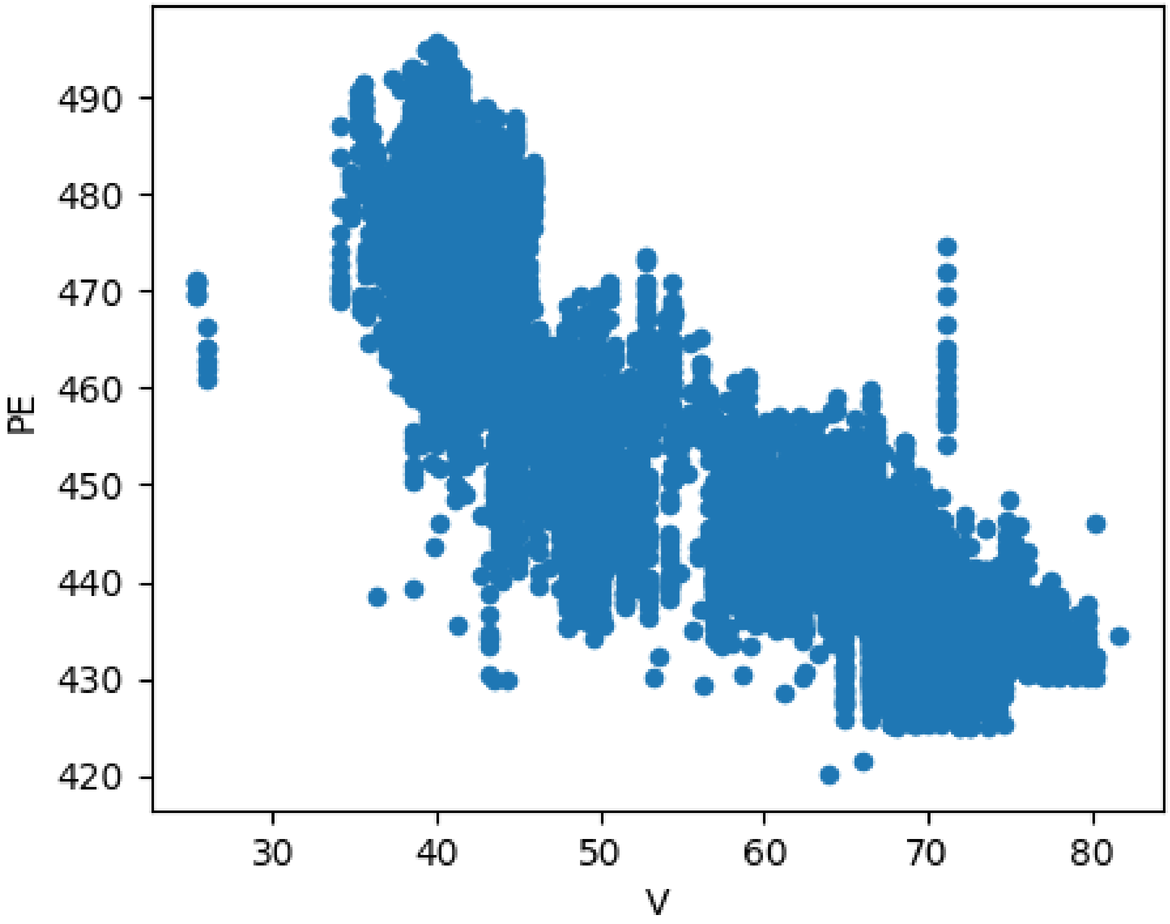

Figure 5: Scatter plot of V vs PE (cleaned dataset).

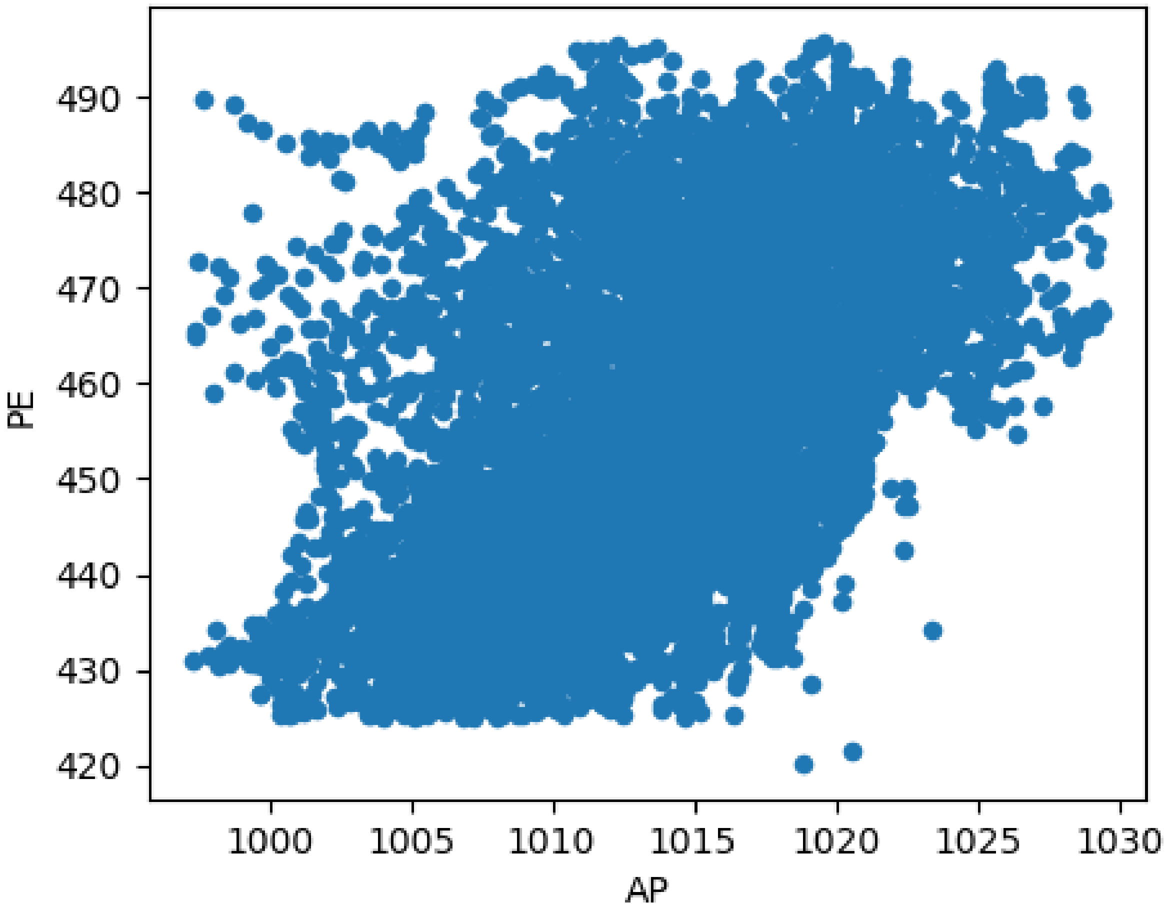

Figure 6: Scatter plot of AP vs PE (cleaned dataset).

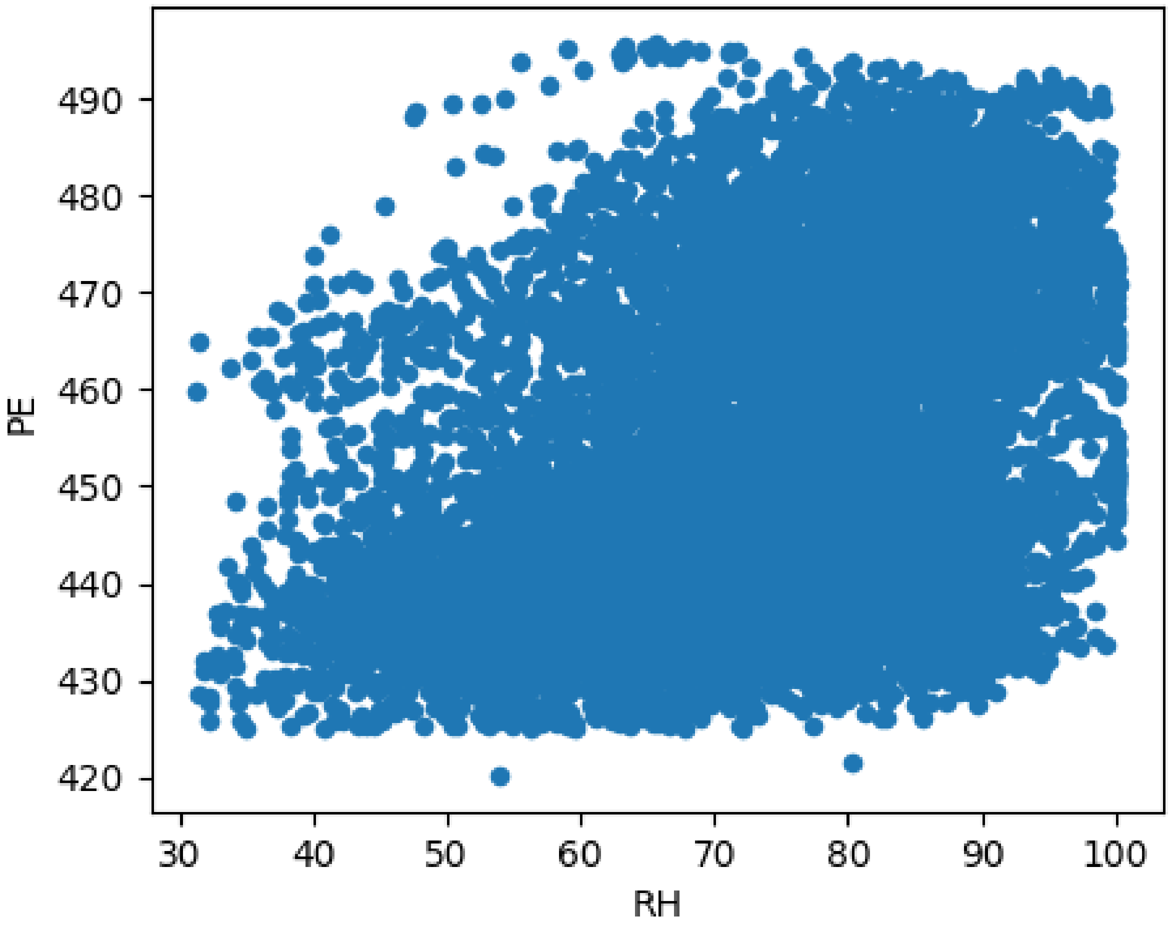

Figure 7: Scatter plot of RH vs PE (cleaned dataset).

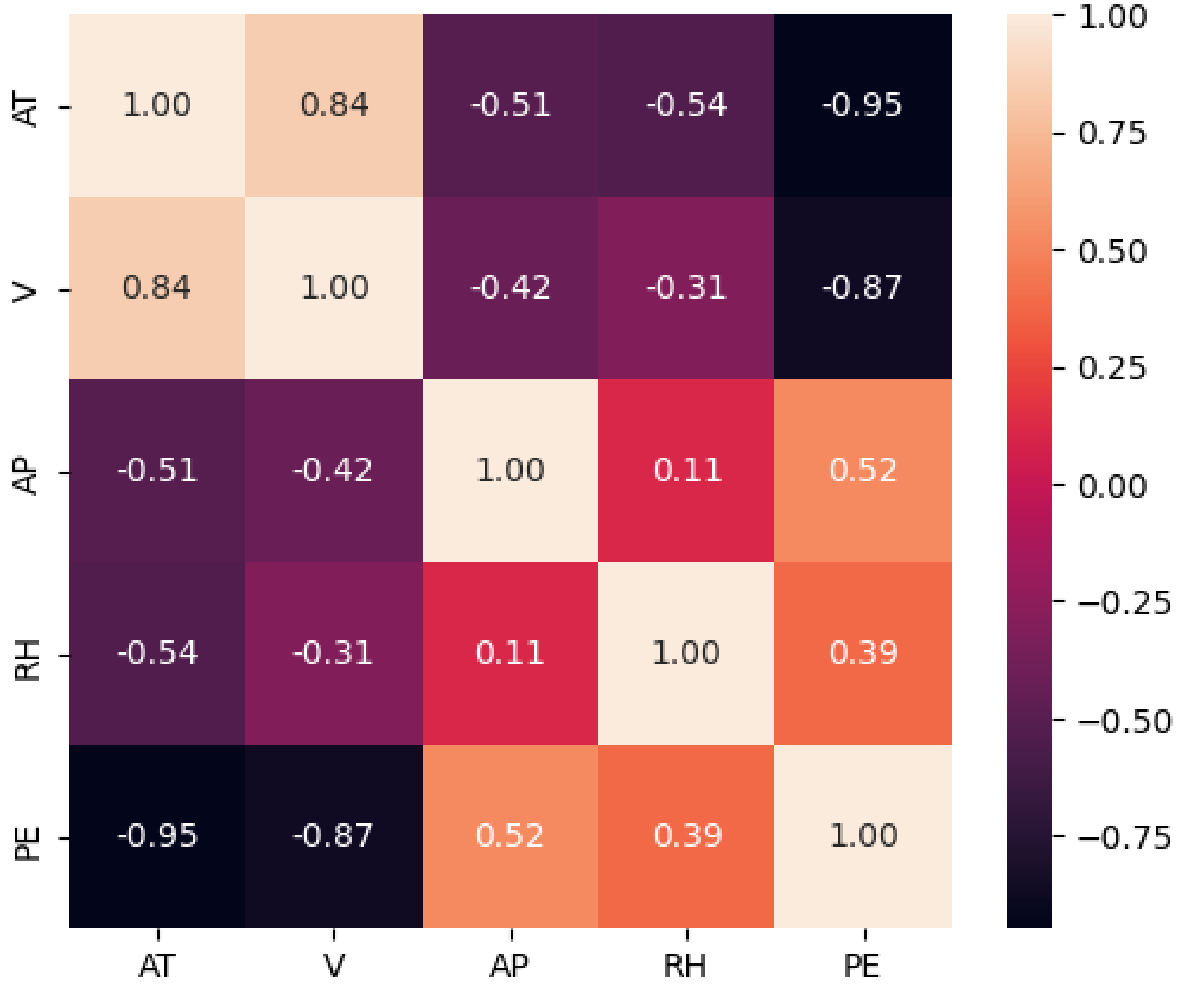

Figure 8: Correlation heatmap for AT, V, AP, RH, and PE (cleaned dataset).

EDA summary

The boxplots (Figures 1-2) were used to detect extreme values across AT, V, AP, RH, and PE, after applying an IQR-based rule, outlier rows were removed to improve robustness. The histograms (Figure 3) show continuous distributions for all variables, with PE spanning approximately 420-496 in the cleaned dataset. Scatter plots (Figures 4-7) indicate strong negative trends for AT and V against PE, while AP and RH show weaker positive relationships. The correlation heatmap (Figure 8) confirms these patterns quantitatively, motivating feature set design that prioritises AT and V, retains AP for additional signal, and treats RH as the weakest predictor.

Task 1 - Class labelling (balanced classes)

Because PE is numerical, it was converted into 4 nearly balanced classes using quantile binning (qcut).

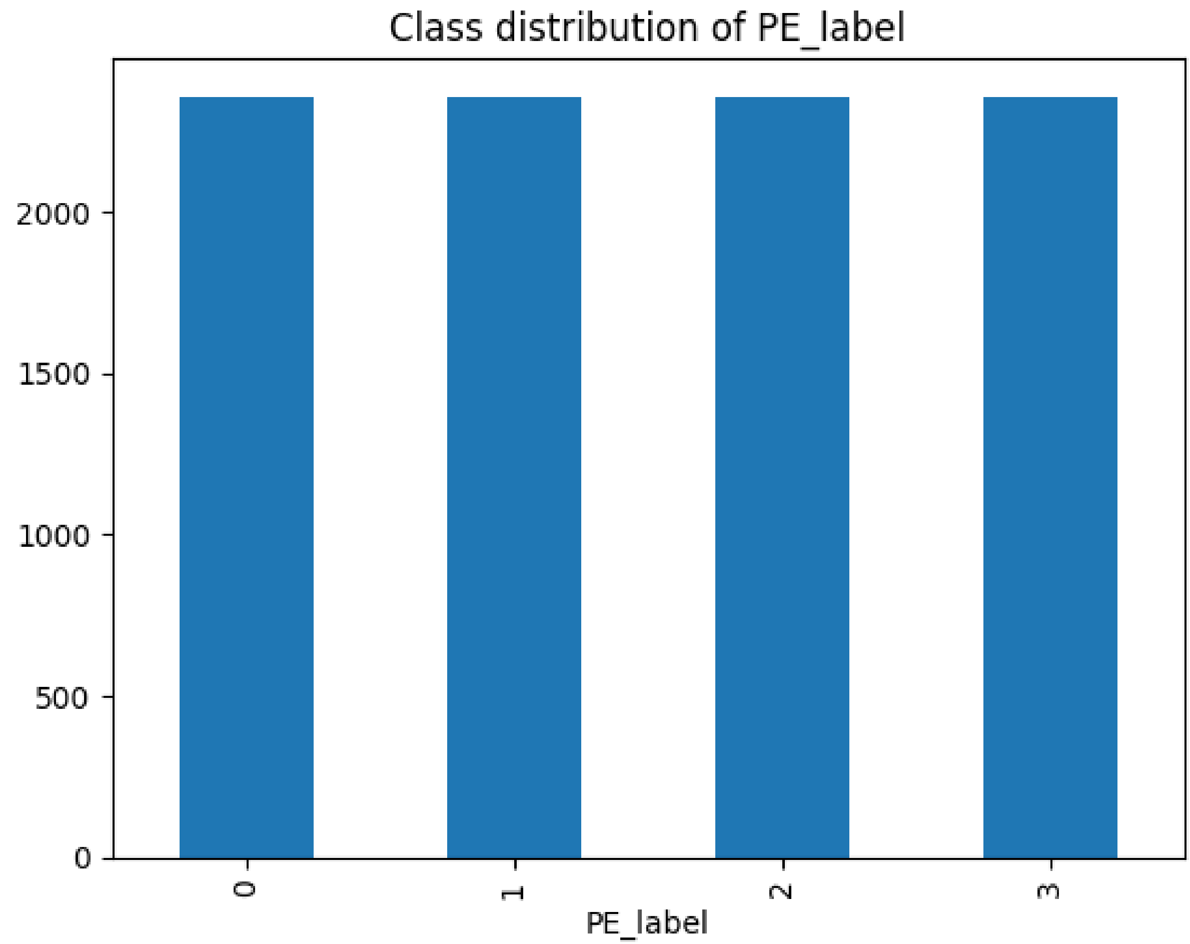

Figure 9: Class distribution of PE_label (4 quantile-based classes).

Class	Count	Proportion	Notes
0	2356	0.250	Quantile bin
1	2356	0.250	Quantile bin
2	2355	0.250	Quantile bin
3	2356	0.250	Quantile bin

Task 2 - Normalisation and integer categorisation

All predictors are numeric. Therefore, normalization was applied using StandardScaler within the modelling pipeline.

The original dataset contained no categorical input features, so integer categorisation/encoding of non-numeric variables was not required for the base dataset.

Task 3 - Feature engineering (EDA-driven)

EDA showed AT and V have the strongest relationships with PE and that the effects may not be perfectly linear across the full range. To capture potential non-linear and interaction effects, squared terms and interaction terms were added and evaluated alongside the original predictors.

Engineered feature	Definition	Motivation
AT2	AT^2	Capture potential non-linear temperature effect
V2	V^2	Capture potential non-linear vacuum effect
ATxV	AT x V	Interaction between the two strongest predictors
APxRH	AP x RH	Interaction between moderate predictors

Task 4 - Decision trees using 5 feature sets

- Model: Decision Tree Classifier
- Train/test split: 80/20 (stratified by class)
- Preprocessing: StandardScaler applied to selected features
- Metrics: Accuracy and Macro F1-score

Feature set	Features included
S1_raw_all	AT, V, AP, RH
S2_drop_weakest	AT, V, AP
S3_top2	AT, V
S4_raw_plus_engineered	AT, V, AP, RH, AT2, V2, ATxV, APxRH
S5_engineered_only	AT2, V2, ATxV, APxRH

Feature set	Features	Accuracy	Macro F1
S2_drop_weakest	3	0.816446	0.817042
S4_raw_plus_engineered	8	0.815915	0.816434
S3_top2	2	0.804244	0.804513
S1_raw_all	4	0.797878	0.797995

S5_engineered_only	4	0.797347	0.797834

Brief observations

The best-performing feature set was S2_drop_weakest (AT, V, AP), which aligns with the EDA finding that RH is the weakest predictor. Adding engineered features (S4) did not materially improve performance, suggesting the decision tree already captures non-linear structure without needing squared/interaction terms to significantly boost accuracy. Using only the top two predictors (S3) reduced performance relative to S2, indicating AP provides additional predictive information.

Conclusion

This report applied an EDA-to-modelling workflow on the CCPP dataset. After removing duplicates and IQR-based outliers, PE was converted into balanced quantile classes and evaluated using decision tree classifiers across five EDA-derived feature sets. The strongest performance was achieved by retaining AT, V, and AP while excluding RH.

Appendix - Code link

Notebook Link