# Supervised and Unsupervised Machine Learning: Heart Disease Prediction and NILM Appliance Clustering

**Course:** IAI4100 Introduction to AI
**Date:** Spring 2025

# Table of contents

# 1. Introduction

This report addresses both the supervised task of prediction/classification and unsupervised tasks of clustering and dimensionality reduction, following the required pipeline steps. The focus is on building models that are accurate yet interpretable, reflecting in challenges such as features selection, evaluation methods and class imbalance. The supervised part applies Logistic Regression on a medical dataset to predict heart disease. The unsupervised part used Keans clustering with and without PCA to group appliance events in household electricity data. Both tasks highlight trade-offs between predictive performance and interpretability. Ethical and sustainability aspects are also considered.

# 2. Supervised Learning: Heart Disease

## 2.1 Dataset and exploration

The dataset used in this study comprises 303 patient records from the Cleveland Heart Disease dataset, which remains a widely cited benchmark in cardiovascular prediction research (Detrano et al., 1989). Each observation contains 13 clinical attributes, covering demographic information (age, sex), cardiovascular measures (cholesterol, resting blood pressure), and functional indicators such as chest pain type, maximum heart rate, and ST depression (oldpeak). The dependent variable is binary, distinguishing between patients with and without a diagnosis of heart disease. The target distribution is relatively balanced, with 160 patients (53%) without heart disease and 143 patients (47%) with heart disease, reducing the risk of class imbalance in training.

Exploratory data analysis (EDA) highlighted several clinically relevant patterns. Gender distribution showed that 206 patients (68%) were male, and 97 patients (32%) were female. Within this, heart disease prevalence was higher among men, where 114 males (55%) were diagnosed positive, compared to 29 females (30%).

Chest pain type also proved to be a strong predictor: patients with asymptomatic chest pain (n = 86) exhibited markedly higher disease prevalence, with more than 75% diagnosed as positive, whereas those reporting typical or nonanginal pain had considerably lower rates.

For continuous predictors, we calculated thresholds based on median values within the disease-positive group. The median maximum heart rate among patients with heart disease was 142 beats per minute, which we used as a reference cutoff. Patients below this threshold were disproportionately classified as positive, suggesting that reduced exercise tolerance is linked to elevated risk. Similarly, analysis of ST depression showed that patients with oldpeak values > 1.5

experienced sharply higher prevalence, with 75% of this subgroup diagnosed with heart disease, compared to 41% in the medium range (0.1–1.5) and 26% in the low-risk group (oldpeak = 0).

These findings confirm that both categorical and continuous variables contribute significantly to prediction. Figure 1 (top left) confirms distribution of heart disease, (top right) disease by gender, (bottom left) chest pain types and (bottom right) age versus max heartrate

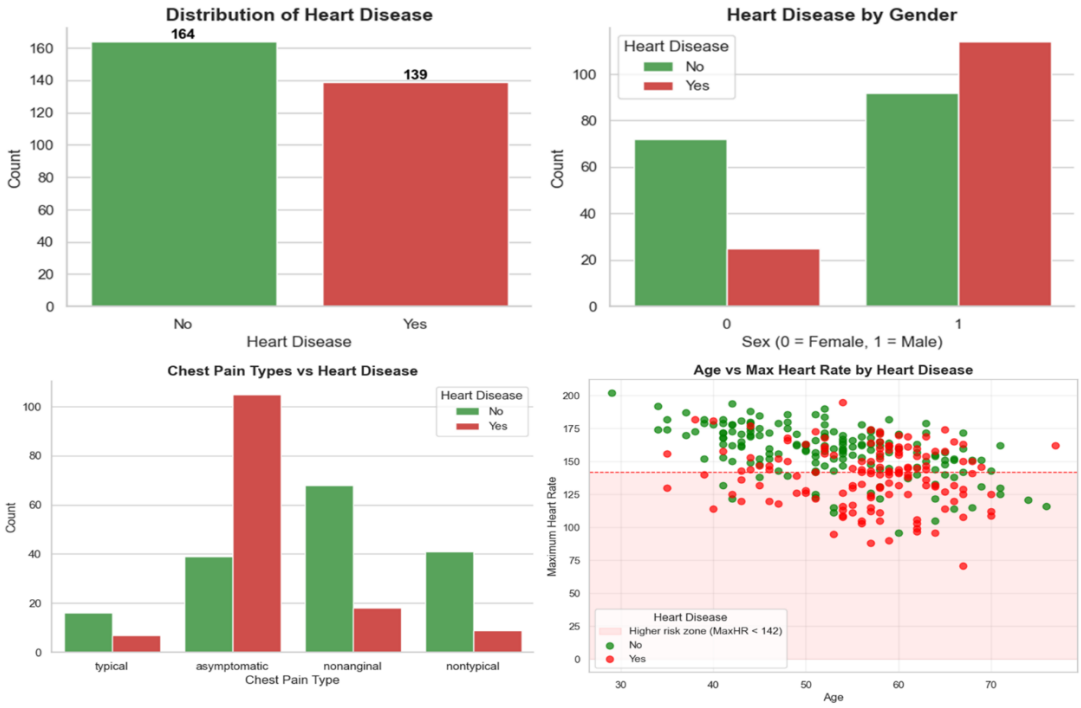



*Figure 1. Exploratory Data Analysis of the Heart Disease Dataset*

To prepare the data for modeling, we implemented structured preprocessing. Five categorical variables were one-hot encoded, increasing the predictor set to 25 features. In addition, certain variables originally stored as numeric but representing categories (RestECG, Slope and Ca) were transformed into dummy variables to ensure they were correctly treated as categorical predictors (James et al., 2013, p. 84). A variance threshold filter was then applied, removing five low-information features (ChestPain_typical, Thal_fixed, RestECG_1, Slope_3, Ca_3.0). As “It is often the case that some or many of the variables used in multiple regression model are in fact not associated with the response” (James et al., 2013, p. 204).

Subsequently, correlation analysis identified three redundant features being >0.85 (Thal_reversable, RestECG_2, Slope_2) were dropped to reduce multicollinearity (James et al., 2013). Following these steps, the dataset was refined to 17 informative predictors, balancing dimensionality reduction with preservation of predictive content. Additional visualization can be seen in Appendix A1-5.

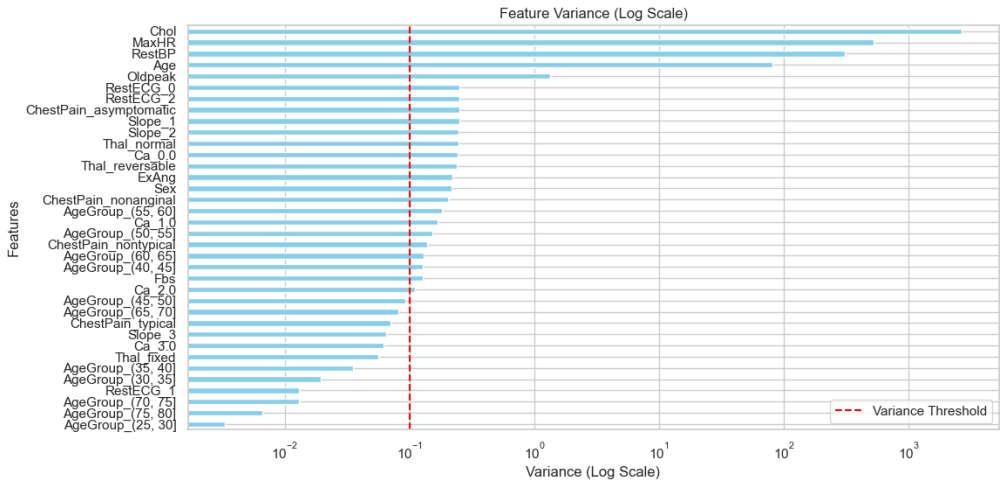



*Figure 2. Variance threshold*

After encoding categorical variables and removing low variance and highly correlated features, the dataset was reduced to 17 informative predictors. These steps ensured valid feature representation, reduced redundancy, and minimized multicollinearity, making the dataset well suited for Logistic regression modelling.

## 2.2 Model selection and methodology

The objective of the study as stated earlier is to predict whether a patient has heart disease which are binary outcomes (0/1). Logistic Regression (LR) suits this prediction task because it was specifically designed for binary classification problems. LR differs from linear regression in that, instead of fitting a straight line to the data, which can give meaningless probabilities below 0 or above 1, LR uses logistic (sigmoid) function to map predictions smoothly into the valid probability range [0,1] (James et al., 2013, pp. 132-133). Mathematically, the model expresses the log-odds of the outcomes as linear functions of the predictors.

*Equation 1 Adapted from (James et al., 2013, p. 132)*

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X.$$

This means each coefficient reflects how a unit change in a predictor multiplies the odds of heart disease. This formulation ensures both valid probability estimates and interpretability. As James explains "if $\beta_1$ is positive then increasing X will be associated with increasing p(X), and if $\beta_1$ is negative then increasing X will be associated with decreasing p(X)" (James et al., 2013, p. 134). For this task, a positive coefficient for feature such as "oldepeak" would indicate that higher

values increase the probability of heart disease, while a negative coefficient for “Cholesterol” would suggest that higher levels are associated wit a lower probability of disease.

Being able to interpret the coefficient is especially important in a medical context, where understanding how predictors affect disease risk and can guide further research. Important to note as LR remains one of the most widely used methods because it models class probabilities directly and transparently (Hastie et al., 2017). Its practical strength is also confirmed by (Mao et al., 2025), who in a systematic review of supervised models on medical datasets reported that LR achieved accuracy score between ~91-93% depending on dataset and setup. In the following section I will evaluate its performance using metrics such as accuracy, F1-score, and ROC curves to assess predictive power and fairness across subgroups.

The dataset was divided into training (80%) and test (20%) sets using stratified sampling to preserve the class proportion. This ensures that both “Yes” and “No” cases are fairly represented in both sets, which avoids selection bias and supports reliable evaluation (James et al., 2013). Standard scaling was then applied to the continues predictors including Age, Cholesterol, RBP and oldpeak, to transform them into comparable ranges with mean 0 and variance 1. In practice, this prevents features with larger numeric scales from dominating those with smaller ranges, improving the stability and compatibility of coefficients in regression models (James et al., 2013, p. 217).

## 2.3 Results and evaluation

The LR model achieved an overall accuracy of 0.87, meaning that 87% of the predictions on the test set were correct. Accuracy is a useful baseline metric that tells us the error rate on how well our model performed under training (James et al., 2013, p. 37). However, accuracy alone can be misleading if one of the classes is more important than the other, so other metrics were used as well.

*Tabel 1. Metrics*

| Metric | Value | Why it’s effective for evaluation |
|---|---|---|
| Accuracy | 0.87 | Overall proportion of correct predictions. |
| Precision | 0.81 | Measures reliability of positive predictions. |
| Recall | 0.93 | Critical in medical context to minimize false negatives. |
| F1-score | 0.87 | Balances precision and recall into one metric. |
| AUC (ROC) | 0.96 | Threshold-independent measure of discrimination. |

The confusion metrics (figure 3, right plot) shows that the model correctly identified 27 out of 33 “No” disease cases and 26 out of 28 “Yes” disease cases, with 6 false positives and 2 false negatives. To capture this trade-off, we examine precision and recall. Precision for “disease” was

0.81, indicating that when the model predicts disease, 81% of those cases correct. Recall for disease was higher at 0.93, meaning the model successfully detected most patients with heart disease. This high recall is particularly important in medical contexts were missing a true positive (false negative) is critical.

Moving on precision and recall for “No disease” cases were 0.93 and 0.82. Telling us the model is more reliable at confirming when a patient is healthy but sometimes mislabels healthy patients as sick. Again, this is better than having false positives. F1-score was calculated to 0.87 for our model. F1-score were used to balance the recall and precision as it computes the two metrics into one, penalizing the imbalance of the metrics (Russell & Norvig, 2022, p. 728).

Roc curve (figure 3, left plot) demonstrated the trade-off between sensitivity and false positive rate at all thresholds. “The receiver operating characteristic (ROC) curve plots the true positive rate against the false positive rate for all possible thresholds. The closer the curve approaches the top-left corner, the better the classifier” (Russell & Norvig, 2022, p. 728). Our model achieved an AUC of 0.96, indicating excellent discriminative ability and suggesting that the model can reliably distinguish between patients with and without the condition.

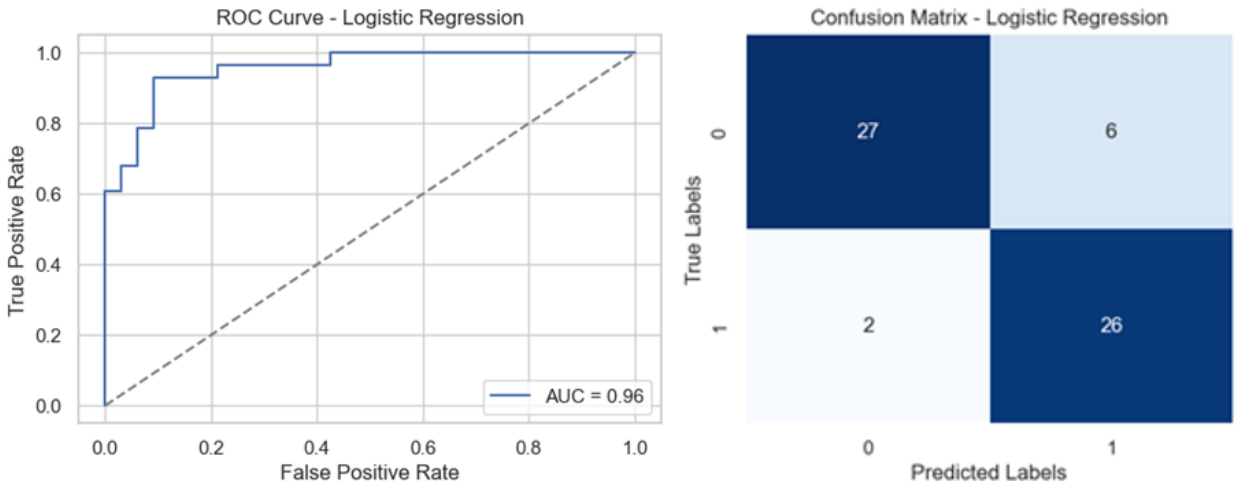



*Figure 3. Roc Curve and Confusion Matrix LR*

## 2.4 Reflection

This project used logistic regression on the heart disease dataset after preprocessing and feature engineering. The model performed well, with 87% accuracy and an AUC of 0.96, showing strong and balanced predictions. Just slightly below the reported accuracies (~91-93%) from (Mao et al., 2025) These results demonstrate that logistic regression is both effective and interpretable for medical risk prediction. However, the model was only tested on one dataset and may generalize differently in other medical contexts. Logistic regression does offer interpretability, more complex models such as ensemble methods might improve accuracy, though at the cost of transparency. Future work could compare different supervised models and evaluate subgroups fairness more explicitly, to ensure robustness in real world healthcare applications.

## 2.5 Ethical measures across the pipeline

All data scientists and engineers face ethical considerations and there are some principles set to follow. The most commonly used are: Ensure safety, Establish accountability, Ensure fairness, Uphold human rights and values, Respect privacy, Promote collaboration, Avoid concentration of power and provide transparency (Russell & Norvig, 2022, p. 1038).

For the data collection part, a well-established, anonymized dataset were chosen, but stewardships over the dataset as a user remains critical. “Data collectors have a moral and legal responsibility to be good stewards of the data they hold” (Russell & Norvig, 2022, p. 1041). For preprocessing the dataset avoids sensitive identifiers, gender is imbalanced. This connects to the fairness, as Russel and Norvig states that “designers of machine learning systems still have moral responsibility to ensure that their systems are in fact fair” (Russell & Norvig, 2022, p. 1043). Including subgroups performance metrics for recall for male vs female patients would make the models fairness evaluation more robust.

Further, the variance threshold and correlation filtering already serve interpretability by simplifying the model. De-biasing the data level is recommended using techniques like SMOTE to generate synthetic data to unbalance the classes. For example, the count of gender in the Cleveland dataset is balanced, but deeper analysis revealed that men had a higher recorded heart disease than women. Making the model performs better for men than women in predicting heart disease.

For modeling and evaluation, accountability means verifying performance under uncertainty. “We need to verify the accuracy and fairness of the results, even if in the face of uncertainly that makes an exact results unknowable” (Russell & Norvig, 2022, p. 1047). The supervised pipeline includes recall and accuracy, but recall is more critical in medical ethics, where false negatives are more dangerous than false positives. Illustrated in figure 4 the confusion matrix. As well with the LR model there is possible to inspect the features to understand why a patient is classified at risk. In other words; “When an AI system turns you down for a loan, you deserve an explanation” (Russell & Norvig, 2022, p. 1048).

Monitoring and auditing require lifecycle process. In practice, meaning to version-control the dataset, preprocessing (variance threshold 0.10, correlation cut off at 0.85), logging experiment results, and documenting model changes for future auditing. In practice, anyone downloading notebooks for this project should be able to run it on their own computer and get the same result.

Finally, in deployment, ethical responsibility includes ensuring that clinicians understand both the strength and limitations of the model, avoiding blind reliance on predictions.

## 2.6 Improving pipeline sustainability

Sustainability requires efficiency and resource-conscious design. Its great if you can create a model prediction in the stock markets that make you 10 dollars on every trade, but not if its computational cost is 20 dollars for each predictions (Russell & Norvig, 2022). They are talking about trade offs other than loss function (Measurement of errors). What makes LR great is its low computational cost and interpretability.

Still there are several improvements that can be made to the pipeline. Currently the pipeline applies plain 5-fold CV on the training data. Switching to StratifiedKFold (k=5=True) would ensure each fold preserves the balance of disease or no disease, aligning with the stratified train/test split (scikit-learn, 2025b). Increasing fairness without extra computational cost.

Next is preprocessing efficiency. Instead of recomputing one-hot encoding, variance filtering, correlation filtering, and scaling in every run. Wrapping these steps into a single Scikit-learn pipeline allows caching and avoids redundant transformations, saving time and resources (scikit-learn, 2025a).

Dimensionality reduction is also a way to make the pipeline more sustainable. Variance and correlation were implemented and reduced the features, however PCA would further reduce them combining features into one or more PCA. Then again remember the tradeoffs mentioned in the earlier example. A doctor won't know what PC1 is, but he will understand Max heart rate.

# 3.0 K-means clustering

K-means clustering is a classic algorithm for grouping data when labels are unknown. It aims to discover "Clusters and cluster centers in a set of unlabeled data" (Hastie et al., 2017, p. 460). The K-means pipeline can be summarized as (1) initialize k cluster center, (2) assign each point to the nearest center, (3) update center based on assignment, (4) repeat until convergence (Hastie et al., 2017).

The object of K-means is to minimize the within cluster variation, so that "Within each cluster the average dissimilarity of the observations from the cluster mean, as defined by points in that cluster, is minimized" (Hastie et al., 2017, p. 509).This is often measured as the straight line distance between two points called Euclidean distance.

K-means is attractive because it is simple, computationally efficient, and works well when clusters are compact and similar in size. It is widely used in practice because of this ease of interpretation. However, it also has important drawbacks. The algorithm is guaranteed to converge, but "the result may represent a suboptimal local minimum" (Hastie et al., 2017, p. 510). This means the results can vary depending on initialization.

The book gives a practical example with gene expressing data, where K-means successfully grouped samples of the same cancer type. In fact, the algorithm, even flagged misdiagnosed two breast cancers placed in different clusters were later confirmed to be melanomas. Yet the same example highlights shortcoming in K-means “does not give a linear ordering of objects within a cluster” and, as the number of clusters changes, “the cluster membership can change in arbitrary ways” (Hastie et al., 2017, p. 514). K-means assumes clusters are compact and of similar size, which may not hold in all datasets. Because of this instability, the authors suggest hierarchical clustering may be preferable in some applications, since it organizes clusters step by step, increasing or decreasing their number in a stale way. The theoretical understanding directly informs the unsupervised NILM task in section 4, where K-means was applied to appliance event clustering.

### 3.1 Dimensionality reduction

Dimensionality reductions mean making data simpler by reducing the number of variables while keeping the main patterns. A common method is Principal Component Analysis (PCA). PCA finds new variables, celled principal components, which are linear combinations of the originals. The first component captures the most variance in the data, meaning it shows the strongest pattern The second captures the most variance, but at a right angle to the first, and so on (Hastie et al., 2017).

This makes PCA useful for compressing data and for visualization, since projecting onto just a few components often keeps most of the structure. The answer is simpler data, less noise, and clearer plots. The main drawbacks are that PCA assumes linear relationship and the components can be hard to interpret because it mixes original variables.

## 4.0 Unsupervised learning NILM project

### 4.1 Problem definition

This project addresses non-intrusive load monitoring (NILM) as an unsupervised learning problem. Recovering appliance level activity from whole-house mains power. The dataset is restricted to house 1 of UK-DALE 2015 where the mains channel (“aggregated”) is aligned to a strict six-second grid. House 1 was selected because it contains the largest number of recorded appliances and the most complete timeline. Other houses in the dataset have gaps or fewer appliances. Short gaps (< 2 min) are filled forward, while longer gaps are set to zero, ensuring continuity. A 5W threshold defines on/off states, applied consistently across mains and sub-metered appliance channels. This preprocessing provides a clean time series suitable for event detection and evaluation.

The unsupervised pipeline followed five steps: (1) Preprocessing (gap filling, thresholding, (2) Event detection (ON/OFF intervals), (3) Feature engineering (Amp_W, Duration_s, Energy_Wh, hour_of_the_day), (4) clustering with K-means, (5) Evaluation using accuracy and macro-F1.

Step changes greater than 50W are extracted and paired into ON/OFF intervals of at most 45 minutes. For each event, features are computed.

Amp_W is the power step size. Duration_s is the interval length. Energy_Wh is consumption during the interval and hour_of_day is the time context. All the preprocessing steps mentioned above follows the recommendation of the article introducing the dataset as well as the ReadME guidelines (Kelly & Knottenbelt, 2015).

The resulting dataset (X_events) represents candidate appliance events in feature space. Histograms confirm heavy tailed distribution of events, with frequently small events with frequently small cycles and fewer large, long-duration events (figure 4). Temporal routines are also visible, with events peaking around morning and evening (figure 5).

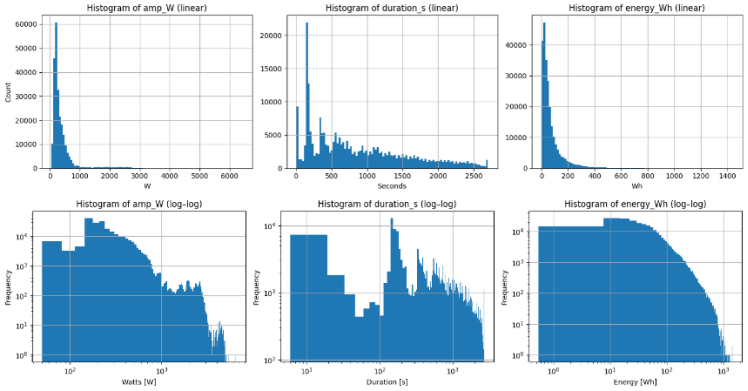



*Figure 4. Event feature distribution (Linear vs. log-scale).*

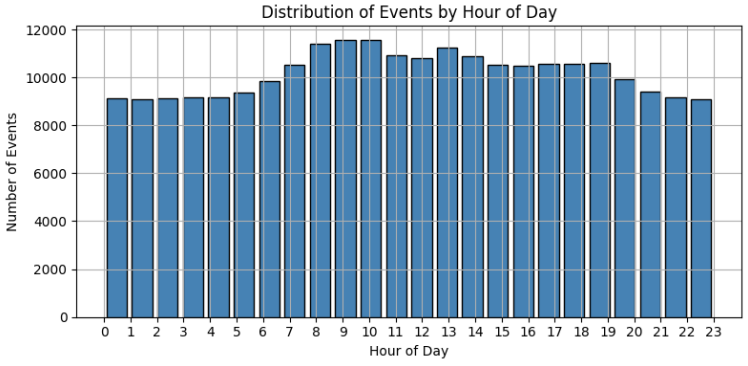



*Figure 5. Distrubution ofevents by hour of day.*

A separate ground truth set of events is built from sub-metered channels. Restricted to five appliances (fridge, kettle, microwave, dishwasher, washing machine). Feature distribution shows fingerprints, particularly in amplitude duration space (figure 6). Important to note, these labels

are used only for evaluation, therefore staying true to unsupervised clustering. Macro F1 was chosen as a metric since it computes F1 for each class separately, then average them equally across classes. Making it more suitable for imbalanced multi-class (Leung, 2022). See appendix B1-9 for more visualizations.

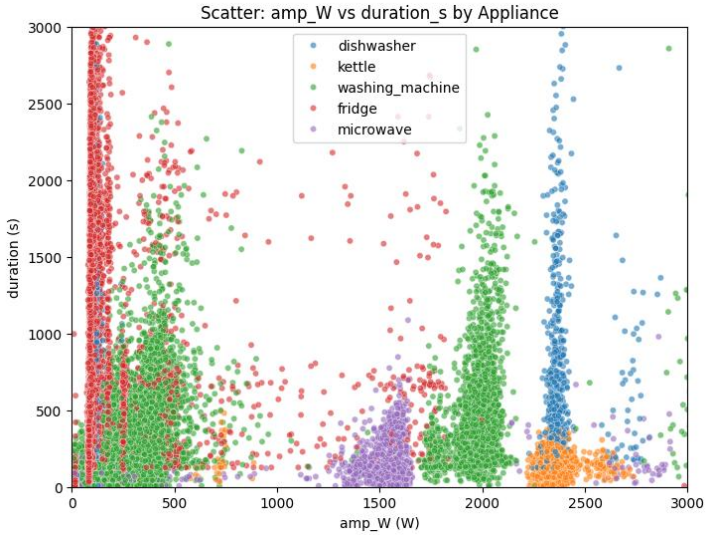



*Figure 6. Appliance fingerprints in amplitude-duration feature space*

## 4.2 The results

Two clustering pipelines were implemented. The baseline pipeline scales the five features with StandardScale and applies KMeans with k=5, chosen to reflect the number of benchmark appliances. Cluster assignments are compared to ground truth through majority voting. Results show accuracy of 0.779 but a much lower macro f1 of 0.286, driven by class imbalance, fridge events dominate and achieve near perfect recall, whereas microwave events are rarely covered. The variant pipeline combines PCA 2 components with K-means. This dimensionality reduction slightly improves balance, yielding macro f1 of 0.302, while accuracy remains unchanged at 0.779. However, PCA improves separation for kettle events but collapses recognition of washing machine, dishwasher and microwave (recall = 0) see figure 7. It shows that while PCA improved balance slightly (macro-F1 0.302), it introduced new weaknesses, underscoring the trade-offs of dimensionality reduction in unsupervised clustering.

*Tabel 2. NILM metrics*

| Pipeline | Accuracy | Macro-F1 | Strengths | Weaknesses |
|---|---|---|---|---|
| K-means baseline | 0.779 | 0.286 | Strong fridge recall | Poor minority appliance detection |
| PCA + K-means (2 PCs) | 0.779 | 0.302 | Better kettle separation | Collapsed washing machine/dishwasher/microwave |

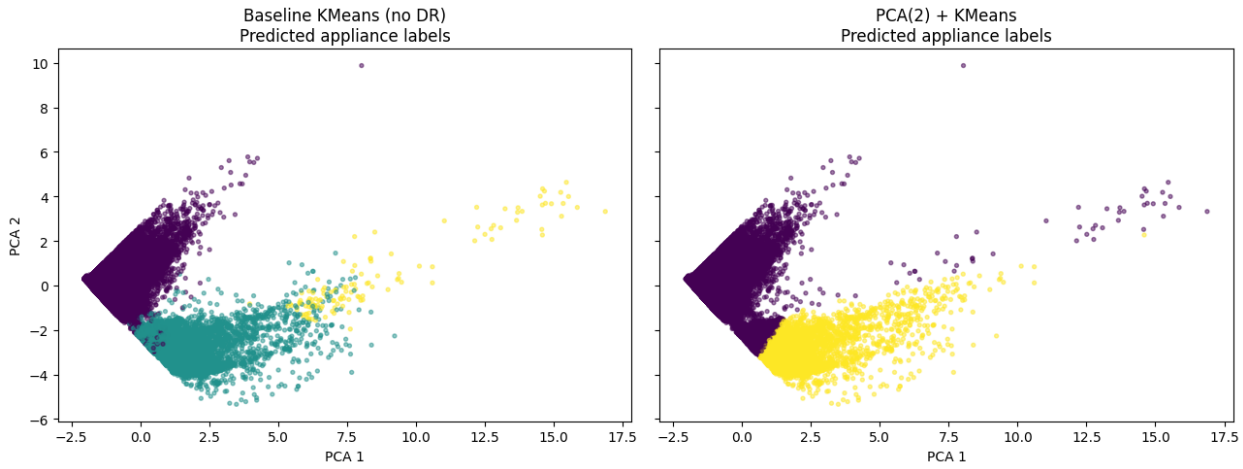



*Figure 7. Comparison of predicted appliance labels using baseline KMeans (left) and PCA(2) + KMeans (right)*

Visualization plays a central role. Scatter plots of Amp_W vs Duration_s illustrate raw cluster boundaries, while PCA projections reveal overlapping distribution between appliances. Confusion matrix highlights systematic misclassifications such as kettle and fridge overlaps. Side-by-side PCA scatter plots compare baseline and PCA pipelines directly, making the trade-offs visible. Dimensionality reduction clarifies clusters for high power short duration appliances but removes subtle variance necessary to distinguish longer, complex cycles.

This design has ethical and sustainability implications. By fixing random seeds and avoiding hyperparameter sweeps, the pipeline limits computational overhead while ensuring reproducibility. Transparency is supported by documenting preprocessing thresholds (6s cadence, 2 min gap rule, 5W on threshold) as the article provided (Kelly & Knottenbelt, 2015). Evaluation explicit reports both accuracy and macro-F1, ensuring the minority appliances are not hidden behind majority performance. These measures align the pipeline with responsible AI principles in terms of reproducibility, fairness and efficiency (Russell & Norvig, 2022).

## 4.3 Pros and cons of dimensional reduction

One clear benefit is visualization and interpretability. As (Russell & Norvig, 2022) emphasize, dimensionality reduction makes complex high dimensional dataset accessible through two dimensional representations. This is a technique used when the four features that were engineered are projected into PCA2. The PCA scatter plots reveal how clusters overlap or separate, offering insights that raw features plots cannot provide. This visualization allows the reader to assess separability of appliances such as kettles versus fridges. similarly (James et al., 2013) argue that reduced presentations can supports improved predictions when the number of predictors is large. In this case, PCA compresses the features space without dramatically lowering classification accuracy, showing that redundancy was reduced while essential structure was preserved.

Another advantage is efficiency. By reducing four features to two, the clustering step becomes computationally lighter. While four dimensions are not extreme, the exercise illustrates dimensionality reduction scales. The exponential growth in grid cells described by (Russell & Norvig, 2022) is managed by projecting into a smaller space, thereby avoiding unnecessary complexity in later clustering and evaluation steps.

However, the results also confirm the tradeoffs identified int the literature. (James et al., 2013) caution that the directions of greatest variance are not necessarily the most predictive. This limitation is visible in the evaluation. PCA improves recognition of some appliances, particularly the kettle, but leads to a collapse in recall for the others, notable the microwave, dishwasher and washing machine. The macro F1 score improves only marginally from 0.286 to 0.302, masking severe drops in minority class performance. This supports the essays warning that dimensionality reduction can bias coefficients and misalign with predictive tasks. Moreover, the pipeline is heavily reliant on manual feature engineering. Events like amplitude, duration, and energy are designed by hand before PCA is applied. Many algorithms only succeed when high dimensional inputs are preprocessed with handcrafted reductions. This dependence risk excluding nonlinear or contextual features that would aid clustering, such frequency harmonics or temporal dependencies. These pros and cons are directly reflected in our results, where PCA improved kettle separability, but reduced performance on more complex appliances.

Several improvements are suggested by the listed literature. First autoencoders could replace PCA as a means of learning presentations directly from the data, capturing nonlinear structure and reducing reliance on handcrafted features (Russell & Norvig, 2022). Second, probabilistic PCA or Gaussian Mixture extensions would model uncertainty explicit, accommodation overlaps between appliances that standard PCA cannot resolve. Third (James et al., 2013) highlights supervised methods such as Partial Least Square that better align dimension reduction with prediction tasks. Incorporating semi-supervised constraints, for example by using a small set of labeled events, could guide clustering more effectively. Finally, class imbalance observed in the pipeline where fridge events dominate, illustrating the curse of dimensionality in practice. Weighting strategies, down sampling, or synthetic minority generation could reduce this distortion before applying PCA.

## 4.4 Conclusion

This report applied supervised and unsupervised learning to medical risk prediction and energy monitoring. Logistic Regression proved both effective and interpretable, while K-means with PCA highlighted the tradeoffs of dimensionality reduction. Ethical and sustainability reflection ensured the pipeline was responsible and reproducible. Overall, the study shows how AI methods can deliver accurate, interpretable, and responsible solutions across domains.

# 5.0 Bibliography


Hastie, T., Tibshirani, R., & Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (Second edition). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 103). Springer New York. https://doi.org/10.1007/978-1-4614-7138-7

Kelly, J., & Knottenbelt, W. (2015). The UK-DALE dataset, domestic appliance-level electricity demand and whole-house demand from five UK homes. *Scientific Data*, *2*(1), 150007. https://doi.org/10.1038/sdata.2015.7

Leung, K. (2022, January 4). Micro, Macro & Weighted Averages of F1 Score, Clearly Explained. *Towards Data Science*. https://towardsdatascience.com/micro-macro-weighted-averages-of-f1-score-clearly-explained-b603420b292f/

Mao, Y., Jimma, B. L., & Mihretie, T. B. (2025). Machine learning algorithms for heart disease diagnosis: A systematic review. *Current Problems in Cardiology*, *50*(8), 103082. https://doi.org/10.1016/j.cpcardiol.2025.103082

Russell, S. J., & Norvig, P. (with Chang, M., Devlin, J., Dragan, A., Forsyth, D., Goodfellow, I., Malik, J., Mansinghka, V., Pearl, J., & Wooldridge, M. J.). (2022). *Artificial intelligence: A modern approach* (Fourth edition, global edition). Pearson.

scikit-learn. (2025a). *Pipeline*. Scikit-Learn. https://scikit-learn/stable/modules/generated/sklearn.pipeline.Pipeline.html

scikit-learn. (2025b). *StratifiedKFold*. Scikit-Learn. https://scikit-learn/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html

# Appendix A. Supervised Project (heart Disease Classification)

Figure A1. Exercise-induced Angina vs Heart disease

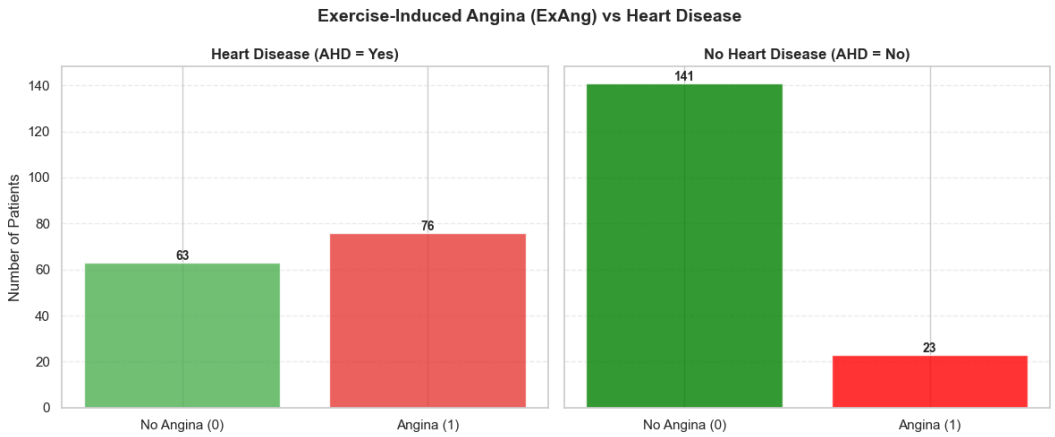



Figure A2. Cholesterol levels by Heart disease (AHD)

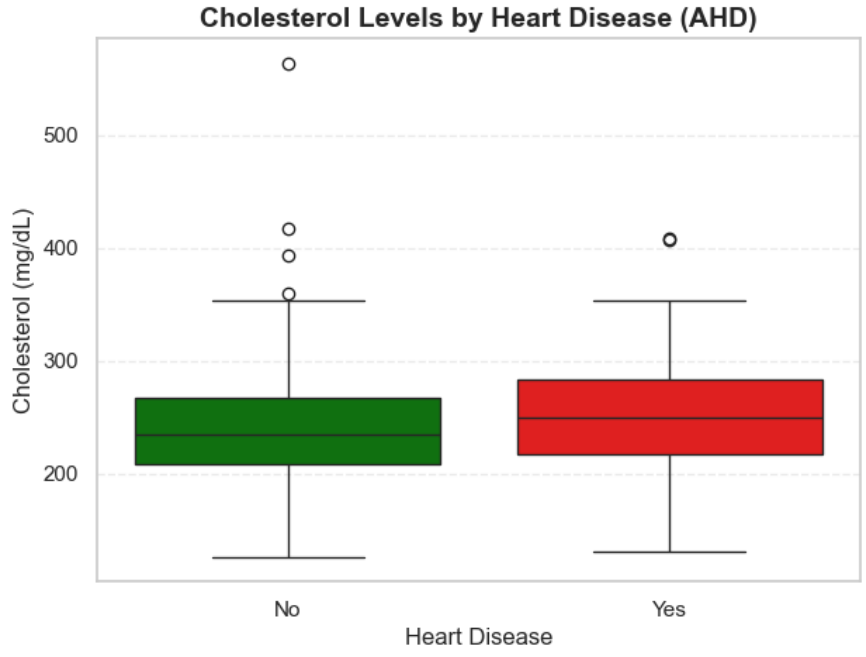


Figure A3. Resting blood pressure by heart disease (AHD)

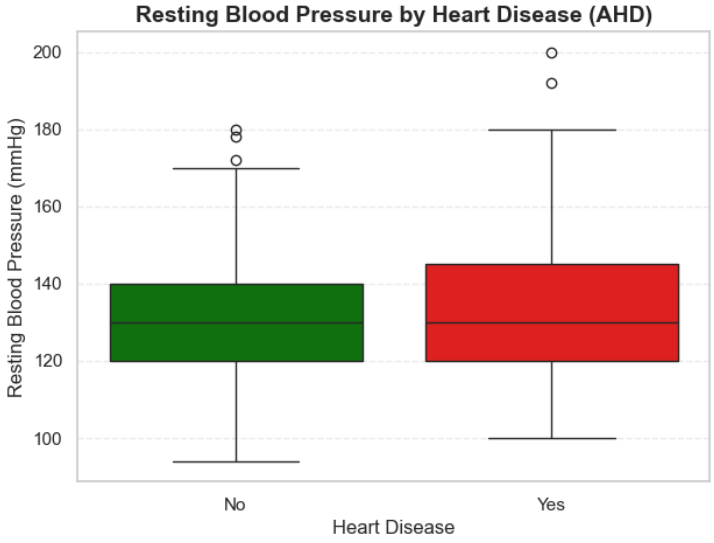



Figure A4. Correlation Heatmap

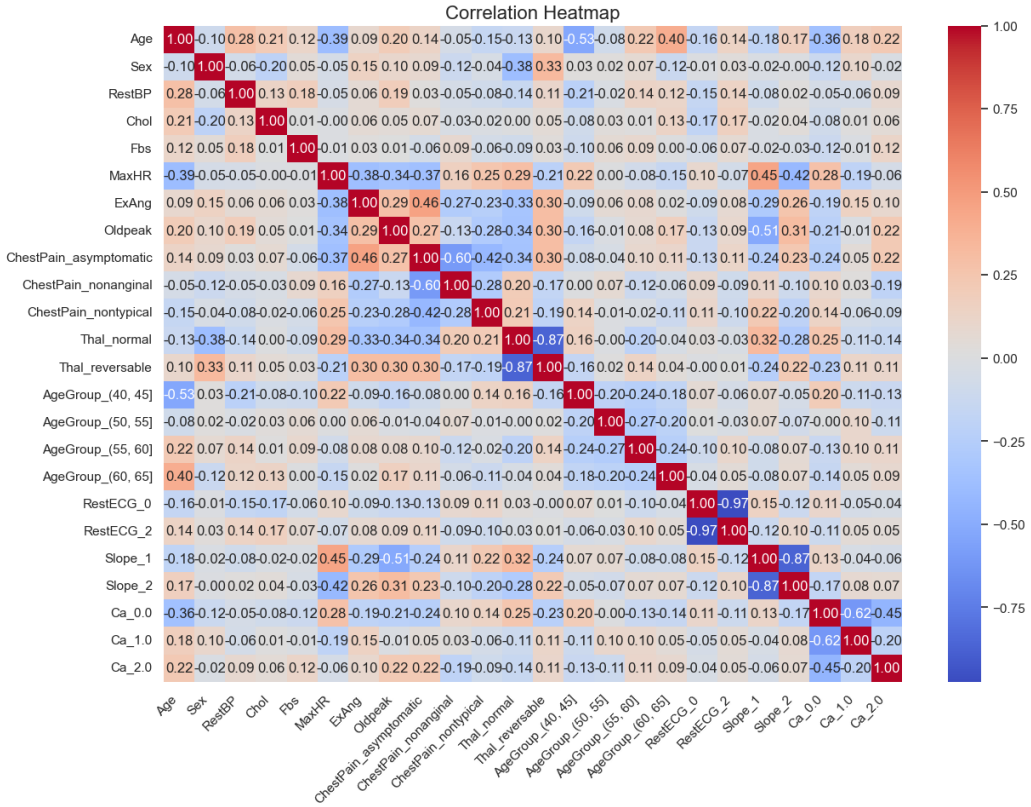
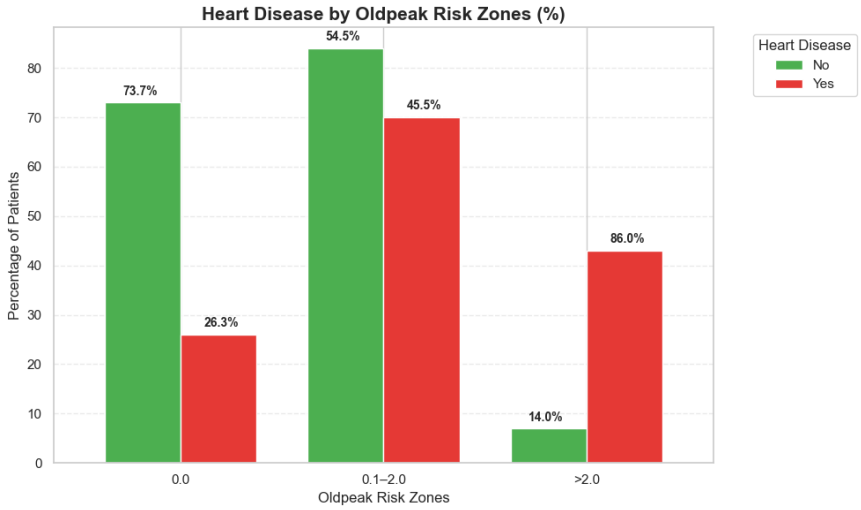


Figure A5. Heart disease by oldpeak risk zones (%)



# Appendix B. Unsupervised Project (NILM clustering)

Figure B1. High/low power appliances

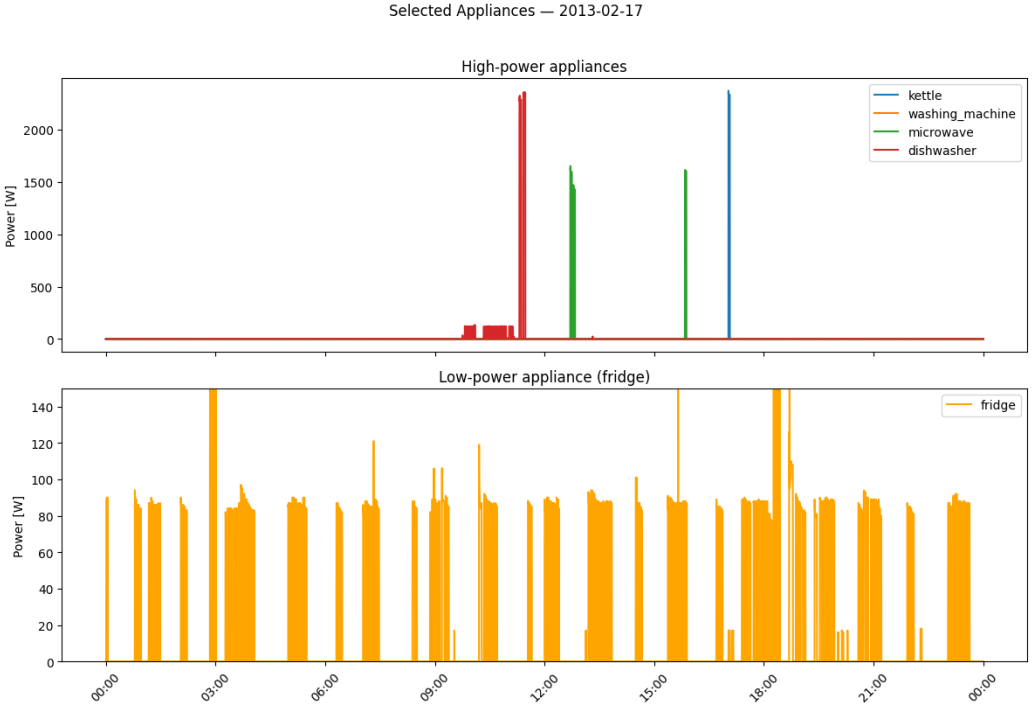



Figure B2. Scatterplot: amp_w vs duration_s

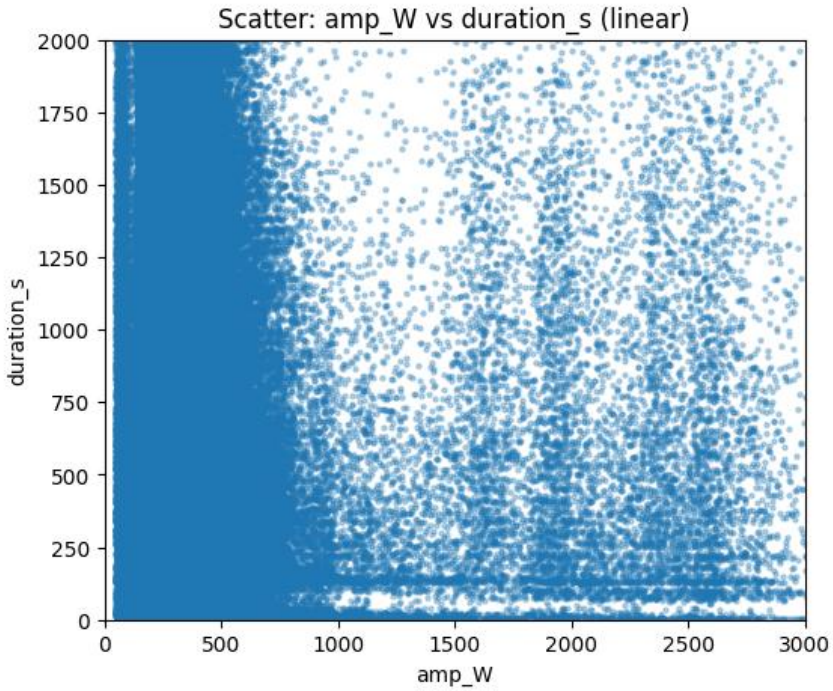


Figure B3. Predicted clusters vs appliances in PCA space

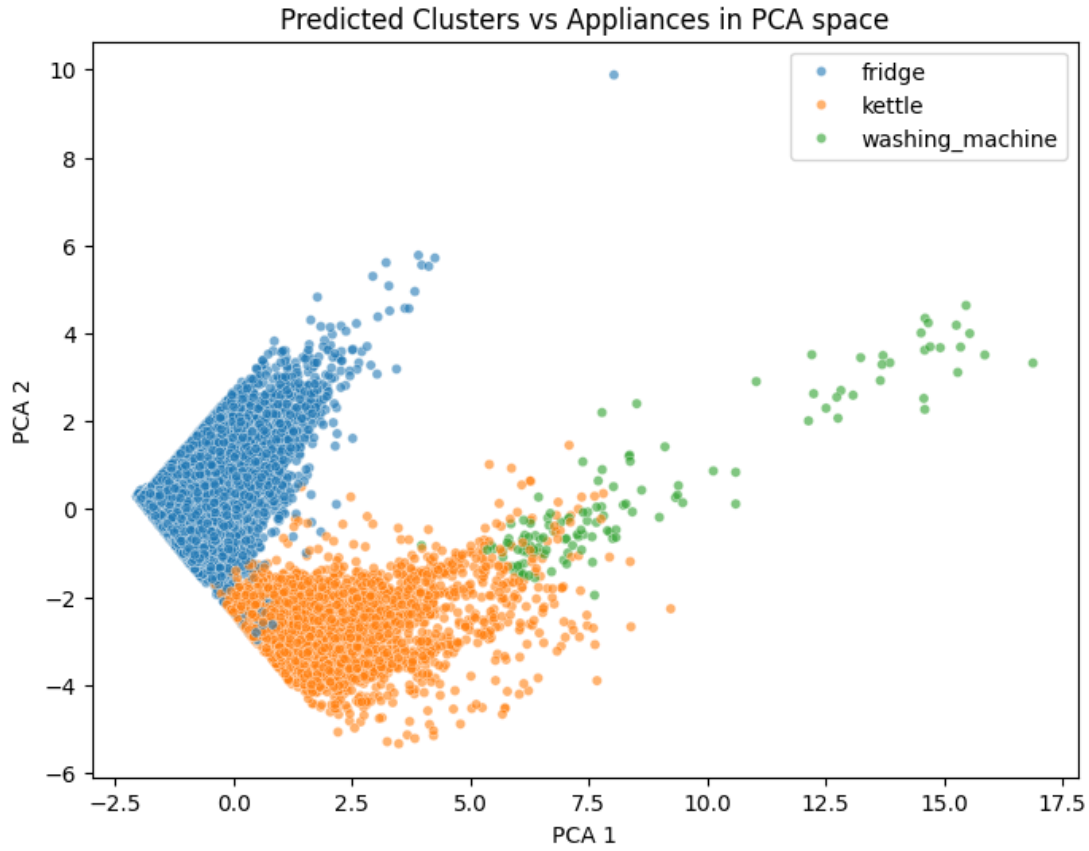



Figure B4. Mains vs sum of appliances

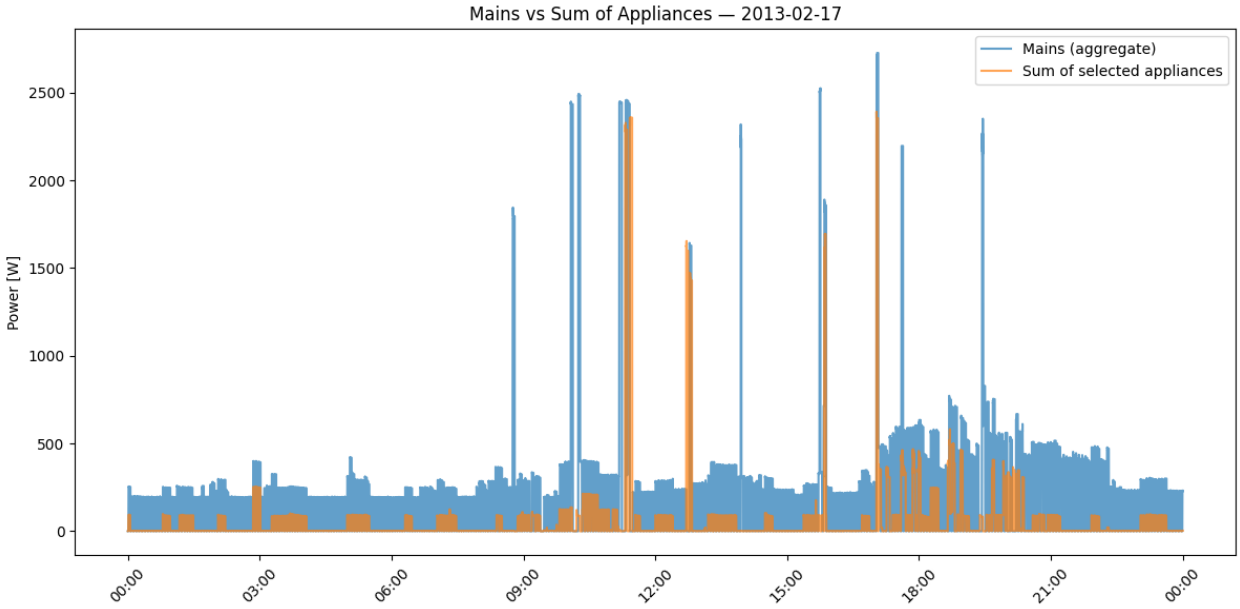


Figure B5. KMeans clusters visualized in PCA space


Figure B6. KMeans clusters on main events

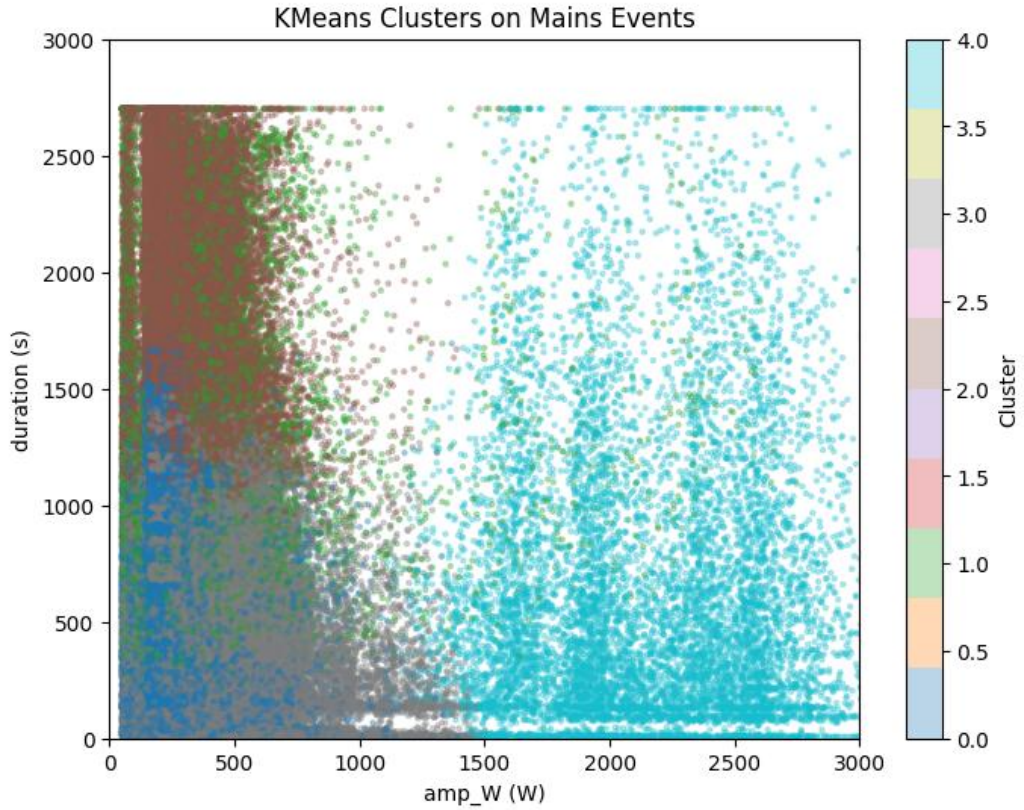


Figure B7. Distrubution of amp_W appliances

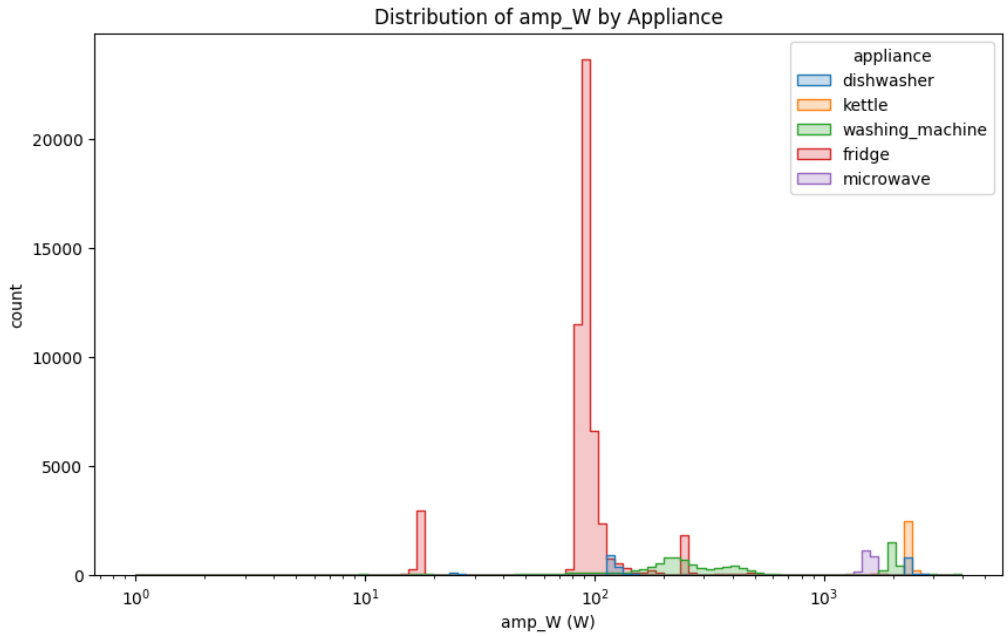



Figure B8. Distrubution of durations_s appliances

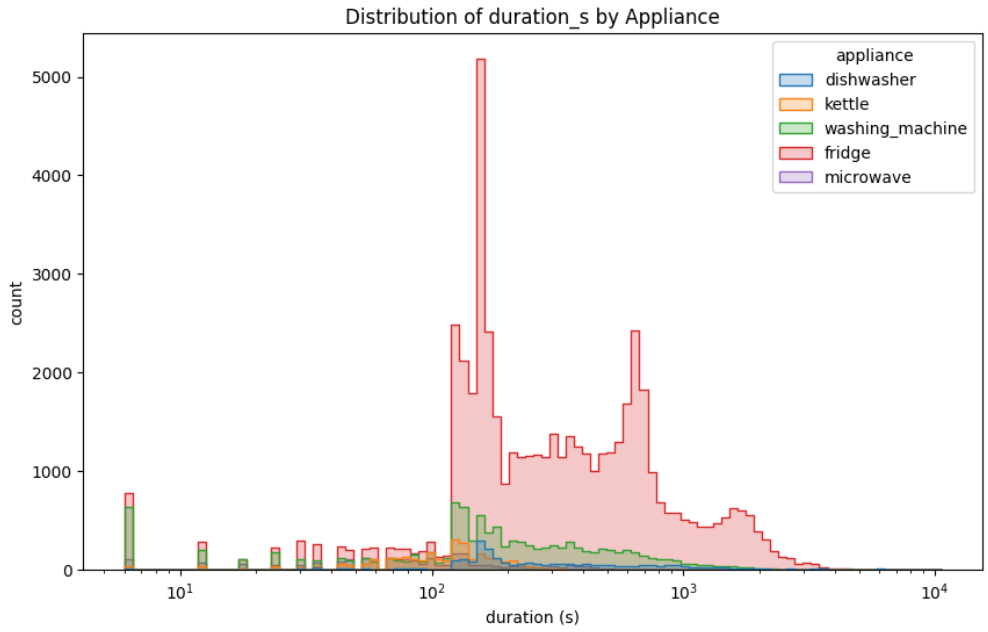


Figure B9. Appliance ground truth in PCA space