Taif University
College of Computers and
Information Technology
Computer Engineering

# Energy Efficiency Prediction using Machine Learning Models

A Course Project Submitted To:
Machine Learning Course
Submitted by:

Shatha Salem Shuraybat
44102547

## Under the Supervision of:

Dr. Nada Al-Tuwairqi

# *Project Report*

## •What is the name of your data?

Energy Efficiency Dataset

## • The source of the data (which database) ?

The data was obtained from the UCI Machine Learning Repository.

## • Link to the original data?

https://archive.ics.uci.edu/ml/datasets/Energy+efficiency

## • Explain the data in words

The dataset includes simulated energy efficiency data for various residential building designs. It has 768 samples, each representing a unique building with different architectural parameters. There are 8 input features like relative compactness, surface area, wall area, roof area, and glazing area. The dataset also includes two target variables: Heating Load (Y1) and Cooling Load (Y2), which quantify the building's energy requirements.

## • Is it a regression or classification problem?

The problem is primarily a regression problem since both targets (Heating and Cooling Loads) are continuous values. However, in one model (Naive Bayes), the regression target was discretized to convert it into a classification task.

## • How many attributes?

There are 10 attributes in total: 8 input features and 2 output variables.

## • How many samples?

There are 768 samples in the dataset.

## • What are the properties of the data ? (statistics)

Heating Load ranges approximately from 6 to 43, with diverse values across all features. For example, Relative Compactness ranges from 0.62 to 0.98. No missing data were observed, and distributions show moderate variance.

## • Are there any missing data? how did you fill in the missing values?

No missing data were found. Therefore, no imputation or filling techniques were applied.

# *Project Report*

## • Visualize the data

We visualized the data using scatter plots (actual vs predicted values) for regression models and confusion matrix for the classification model (Naive Bayes). The ANN model also included a training loss curve.

## • Did you normalize or standardize any of your data? why?

Yes. Standardization was applied before training the SVM and ANN models. These models are sensitive to the scale of features and require normalization for optimal convergence and performance.

## • What type of preprocessing did you apply to your data? List everything and explain why.

- Renamed the dataset columns to meaningful names
- Split data into training and testing sets
- Standardized the input features for ANN and SVM
- Binned continuous labels into categories for Naive Bayes

## • How did you divide the train and test data? what are the proportions?

We used an 80/20 split using train_test_split() from scikit-learn, with a fixed random state to ensure reproducibility.

## •Apply all ML models and report results. What is the best/worst model? Why?

We applied Linear Regression, Decision Tree, Random Forest, KNN, SVM, ANN, and Naive Bayes.
Best: ANN (lowest RMSE, highest $R^2$)
Worst: Naive Bayes (requires classification, not optimal for regression).

# Project Report

- The accuracy of all models using tables and figures?

| Type | Value | Metric Used | Model |
|---|---|---|---|
| Regression | ~5.43 | RMSE | Linear Regression |
| Regression | ~0.0–1.0 | RMSE | Decision Tree |
| Regression | Very Low | RMSE | Random Forest |
| Regression | Moderate | RMSE | KNN |
| Regression | ~2–4 | RMSE | SVM |
| Regression | Lowest | RMSE | ANN |
| Classification | ~60–70% | Accuracy | Naive Bayes |

- Bonus Visualization

We included:
- Scatter plots (actual vs predicted)
- Confusion matrix (Naive Bayes)
- Training loss curve (ANN)

- Explain

This dataset was chosen due to its real-world relevance in sustainable building design. With
global energy demands rising and environmental regulations tightening, optimizing energy
efficiency in building structures is a critical challenge. The dataset simulates various designparameters that affect heating and cooling loads, allowing the application of machine
learning techniques to predict energy needs. Such predictions can guide architects and engineers to make energy-conscious decisions early in the design process, saving costs and
reducing environmental impact. Moreover, applying diverse models on the dataset provided
a hands-on comparison of regression techniques. It helped in understanding the strengths
of tree-based models, the impact of scaling on algorithms like SVM and ANN, and the limitations of classification methods like Naive Bayes for regression tasks. Overall, this project highlights the value of ML in practical applications, especially those related to energy consumption, which is a global priority.

# *Project Report*

Link to your code and data?

https://github.com/shathasalem2003/ML-Project.git