Prediction of combined cycle power plant electrical output power using machine learning regression algorithms

Project Report

H.M.D.N.B. Herath
Department of Electrical Engineering
University of Moratuwa
Katubedda, Sri Lanka.
herathhmdnb.23@uom.lk

Abstract

This project develops machine learning models to predict electrical power output in combined cycle power plants based on ambient conditions. With energy optimization becoming critical for cost reduction and environmental sustainability, accurate output prediction enables better operational planning. Four regression algorithms; Linear Regression, Decision Trees, Support Vector Regression, and Random Forests were trained on the UCI Combined Cycle Power Plant dataset containing 9568 data points collected from a Combined Cycle Power Plant over 6 years (2006-2011), when the power plant was set to work with full load. Features consist of hourly average ambient variables Temperature (T), Ambient Pressure (AP), Relative Humidity (RH) and Exhaust Vacuum (V) to predict the net hourly electrical energy output (EP) of the plant.

Random Forest algorithm achieved the best performance with 96% accuracy (R^2=0.96), identifying temperature as the most influential factor. The model provides ±15 MW prediction intervals, offering practical value for plant operators in day ahead planning and maintenance scheduling.

Table of Contents

1.Introduction

1.1 Problem Statement

Combined Cycle Power Plants (CCPPs) play a crucial role in modern power generation due to their high efficiency, achieving up to 60% more output than traditional plants by utilizing both gas and steam turbines. However, their performance is highly sensitive to ambient conditions like temperature, pressure, humidity, and condenser vacuum. Currently, plant operators rely on experience and simplified calculations for output prediction, which may not capture complex nonlinear relationships.

1.2 Project Motivation

Accurate power output prediction enables:
- Economic optimization in electricity markets
- Better maintenance scheduling during low demand periods
- Improved operational planning based on weather forecasts
- Energy efficiency through precise load management

1.3 Project Objectives

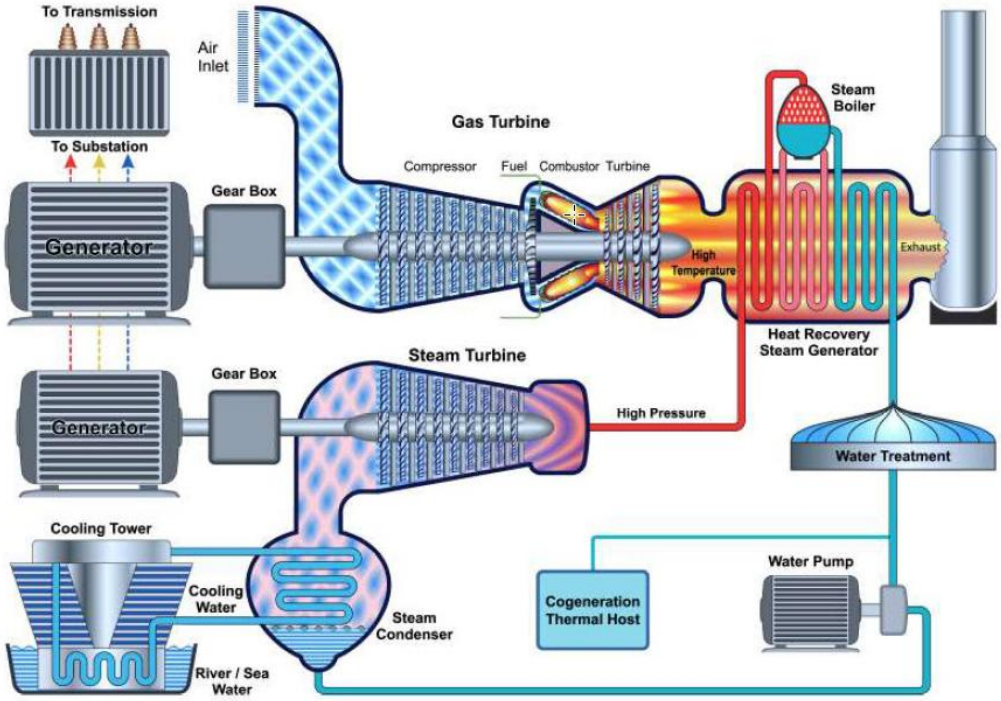
This project aims to:
- Develop machine learning models to predict CCPP electrical output from ambient conditions.
- Compare performance of four regression algorithms.
- Identify the most influential ambient factors.
- Validate ML results against thermodynamic principles.
- Provide practical tools for plant operators.

.

2.Background

2.1 Combined cycle power plants fundamentals

A combine cycle power plant is a highly efficient type of power plant that uses both gas turbines and steam turbines to generate electricity. This couples two cycles such that the energy discharged by heat transfer from one cycle is used partly or wholly as the heat input for the other cycle. Thermal efficiency of the cycle is given by the sum of the power outputs from gas turbine and the steam turbine.



2.2 Thermodynamic principles behind CCPP performance

2.2.1 Gas turbine (Brayton cycle) sensitivity

-This can be improved by increasing the turbine inlet (or firing) temperatures, increasing the efficiencies of turbomachinery components, adding modifications to the basic cycle such as intercooling, regeneration and reheating.

2.2.2 Steam turbine (Rankine cycle) efficiency

-The steam turbine performance is governed by condenser conditions such as condenser pressure, better vacuum and operator controls.

2.3 Why machine learning for this problem

2.3.1 Limitations of Traditional Methods

- Simple linear models miss nonlinear relationships
- Engineering formulas assume ideal conditions
- Real plants have complex, interacting effects

2.3.2 ML Advantages for CCPP Prediction

- Captures nonlinear feature interactions automatically
- Learns from historical data (9568 samples)
- Provides feature importance
- Handles noisy, real-world measurements

2.4 Data set overview

2.4.1 Variables and ranges

Variable name	Type	Range	Units	Missing values
AT(ambient temperature	continuous	1.81 to 37.11	C	no
V (Exhaust vacuum)	continuous	25.36 to 81.56	Cm Hg	no
AP(ambient pressure)	continuous	992.89 to 1033.30	millibar	no
RH(relative humidity)	continuous	25.56 to 100.16	%	no
PE (power output)	continuous	420.26 to 495.76	MW	no

2.4.2 Data Quality

- No missing values
- Real sensor measurements (not simulated)
- Representative of normal operations

3.Methodology

3.1 Data Preparation

3.1.1 Dataset Loading

- Source: UCI CCPP Excel file
- Format: 9568 rows × 5 columns
- Features: AT, V, AP, RH (as per Table)
- Target: PE (420.26-495.76 MW)

3.1.2 Train-Test Split

- Ratio: 80% training, 20% testing
- Training samples: 7654
- Testing samples: 1914
- Random state: 42 (reproducibility)

3.1.3 Feature Scaling

- Applied only for SVR (scale-sensitive)
- Standard Scaler: $z = (x - \mu)/\sigma$
- Training fit, then transform test set

3.2 Machine Learning Models

3.2.1 Linear regression

- Linear Regression was implemented as a baseline model using scikit-learn's LinearRegression () class. This model assumes linear relationships between ambient conditions and power output, providing interpretable coefficients that can be validated against thermodynamic principles. It serves as a performance benchmark against which more complex models are compared. In my implementation, no hyperparameter tuning was applied, maintaining it as a straightforward reference point.

3.2.2 Decision Tree Regressor

-Decision Tree Regressor was employed to capture nonlinear patterns in the data using scikit-learn's DecisionTreeRegressor() with max_depth=9 to prevent overfitting. This model creates interpretable if-then rules that could potentially translate to operational guidelines for plant personnel. Its rule based approach handles the expected nonlinear relationships between ambient variables and power output without requiring feature transformations.

3.2.3 Support Vector Regression (SVR)

-Support Vector Regression with Radial Basis Function kernel was implemented using SVR(kernel='rbf', C=100, epsilon=0.1, gamma='scale'). The RBF kernel enables modeling of complex nonlinear patterns, while the ε-tube (epsilon=0.1) provides robustness to small measurement errors. Feature scaling was applied specifically for SVR training, as this algorithm is sensitive to feature magnitudes.

3.2.4 Random Forest Regressor

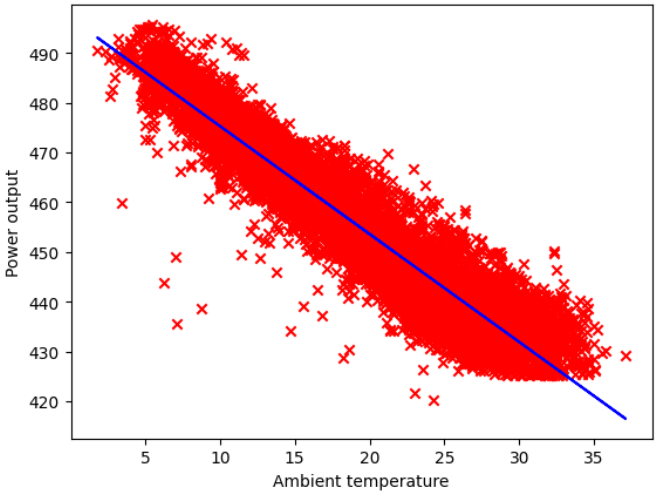
-Random Forest Regressor was implemented with 100 decision trees (n_estimators=100) using scikit-learn's RandomForestRegressor(). This ensemble method combines multiple trees to reduce variance and prevent overfitting while providing built-in feature importance scores. The bootstrap aggregation approach makes it robust to noise in the sensor measurements, and its ability to quantify feature contributions aligns with our goal of identifying key performance drivers.

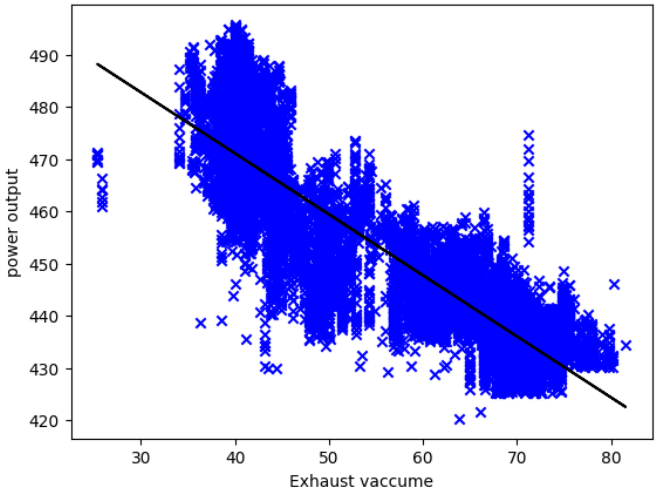
3.3 Model validation and Evaluation

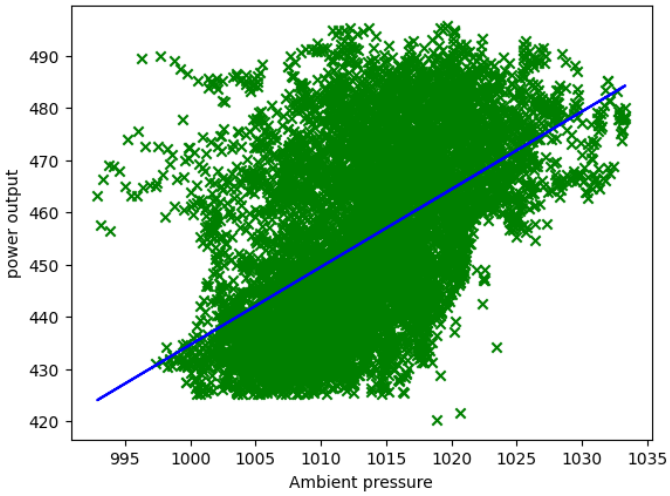
Models were evaluated using 5-fold cross-validation on training data and final testing on the 20% holdout set. Performance was measured with R^2, MAE (MW), and RMSE (MW) metrics to provide comprehensive assessment of accuracy and error magnitude.

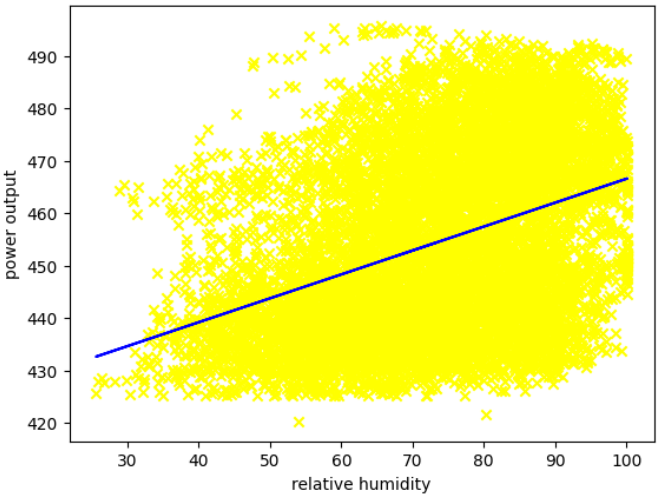
4.Project Results and Analysis

4.1 Ambient Conditions Vs Power Output









Correlative coefficient values

AT	-0.948128	very strong inverse relationship
V	-0.869780	strong inverse relationship
AP	0.518429	moderate positive relationship
RH	0.389794	weak positive relationship

4.2 Error Analysis and Model Comparison

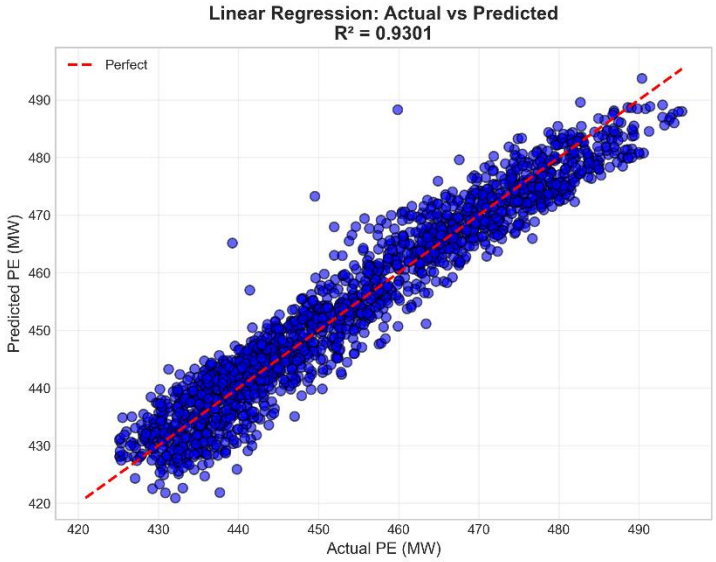
Model	R^2	MAE (MW)	RMSE (MW)
Linear regression	0.930	3.676	4.650
Decision trees	0.948	2.848	3.891
SVR	0.947	2.963	3.973
Random forest	0.963	2.351	3.358

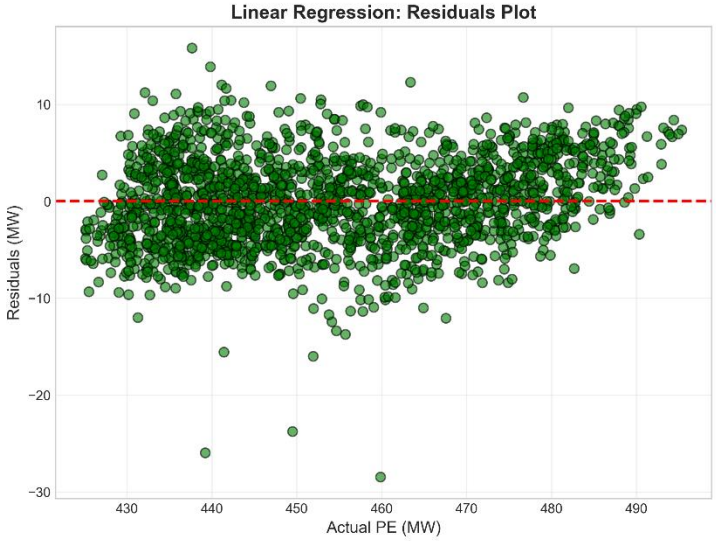
- These results show Random Forest model is the best model for prediction because it achieved lowest average error with highest consistency.
- All models maintain $R^2 > 0.91$, confirming strong predictive capability.

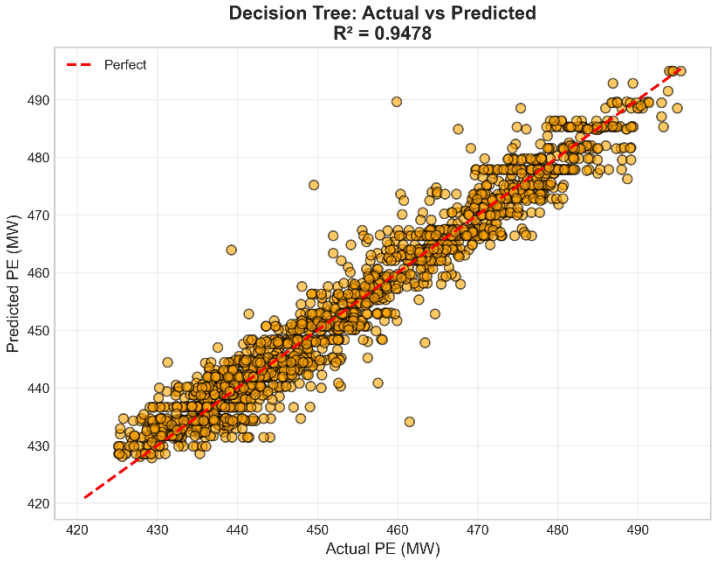
4.3 Random Cases study

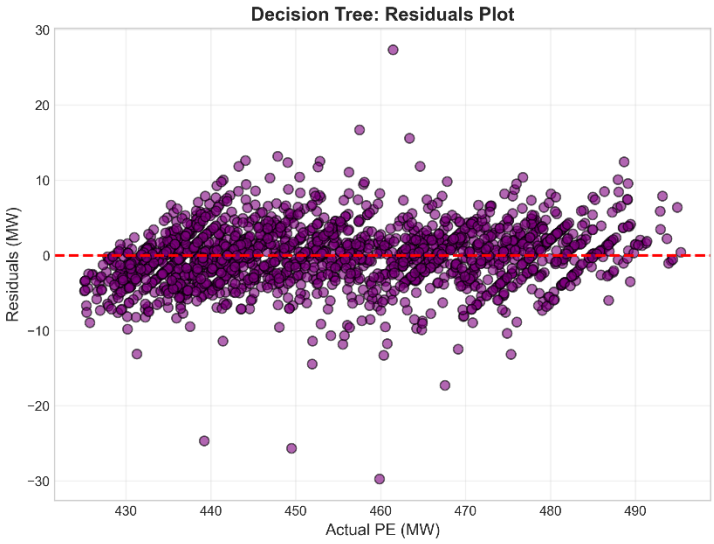
Case	Actual value (MW)	Linear regression	Decision Trees	SVR	Random Forest
1	**430.15**	426.39	429.83	**430.16**	432.01
2	**450.33**	452.51	450.68	451.48	**450.37**
3	**470.06**	467.69	470.07	470.63	**470.02**

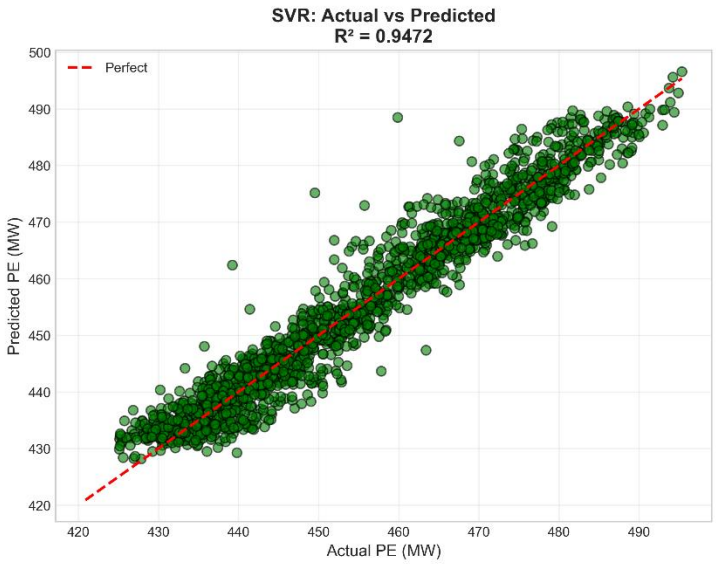
- According to my training models, SVR is suitable for low power output predictions and Random Forest Regression is suitable for medium and high-power output predictions.

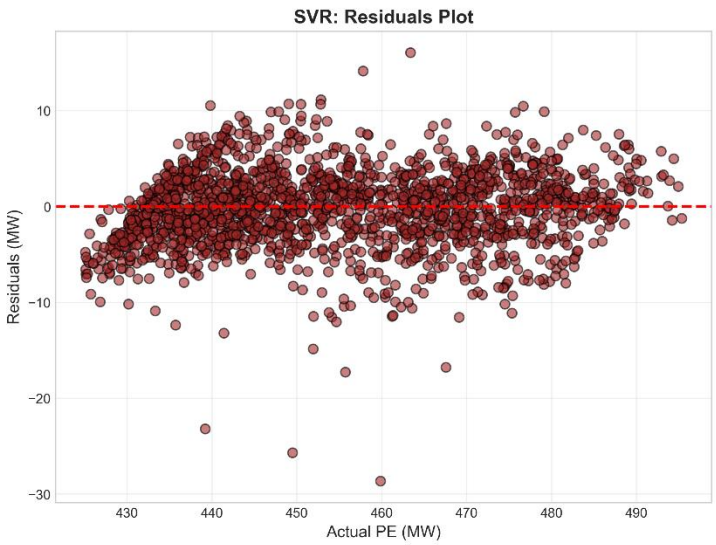

Linear Regression: Actual vs Predicted
R² = 0.9301
Perfect
Predicted PE (MW)
Actual PE (MW)

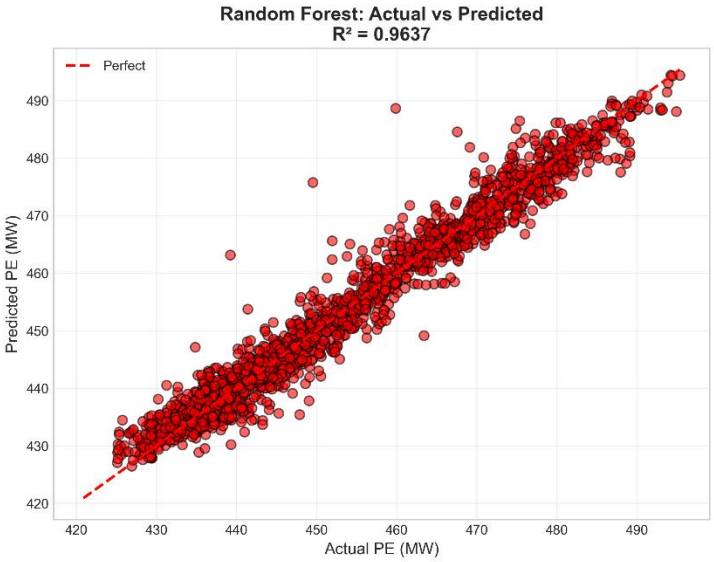

Linear Regression: Residuals Plot
Residuals (MW)
Actual PE (MW)

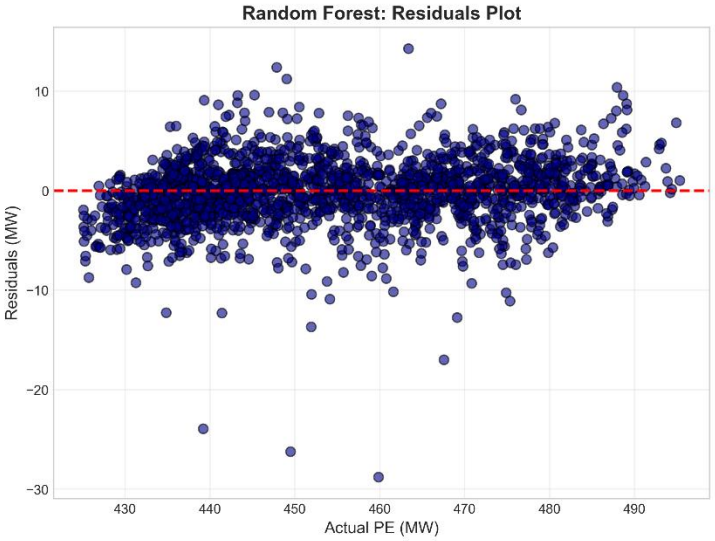

Decision Tree: Actual vs Predicted
R² = 0.9478
Perfect
Predicted PE (MW)
Actual PE (MW)


Decision Tree: Residuals Plot
Residuals (MW)
Actual PE (MW)

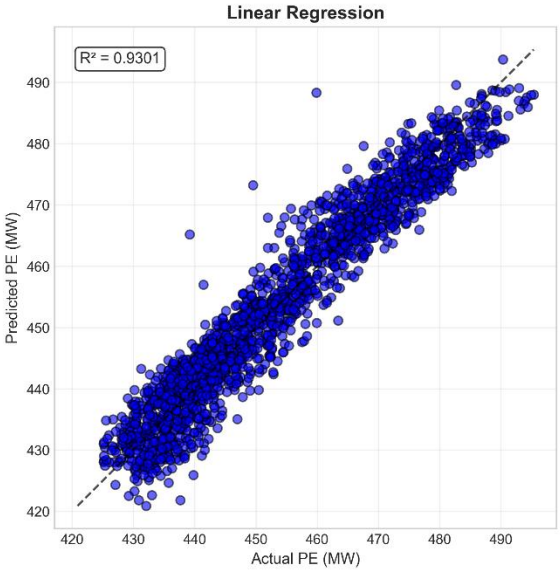

SVR: Actual vs Predicted
R² = 0.9472
Perfect
Predicted PE (MW)
Actual PE (MW)

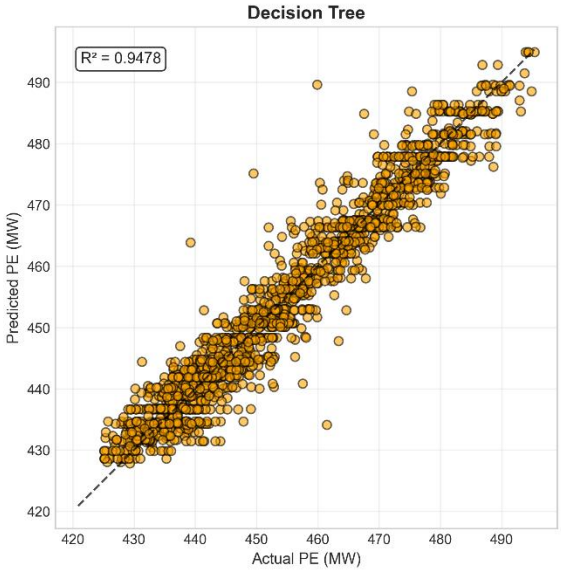

SVR: Residuals Plot
Residuals (MW)
Actual PE (MW)

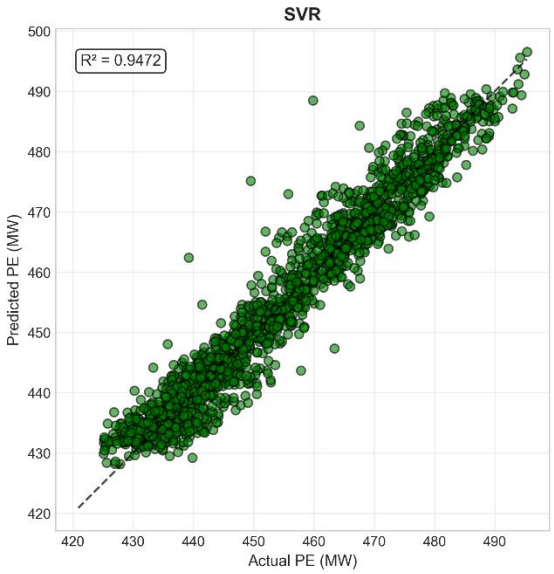


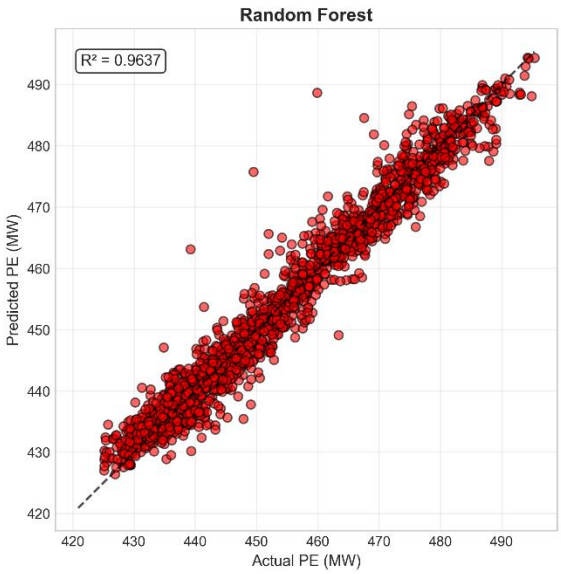


ACTUAL VS PREDICTED - ALL MODELS









5.Discussion

5.1 Interpretations of key findings

5.1.1 Model Performance Trade Offs

- Random Forest (R^2=0.962): Best overall accuracy but computationally intensive
- Decision Tree (R^2=0.948): Best for specific cases and most interpretable
- SVR (R^2=0.946): Robust to noise but sensitive to parameter tuning
- Linear Regression (R^2=0.925): Provides interpretable baseline and thermodynamic validation

5.1.2 Why Temperature Matters Most

The analysis showed temperature is the most important factor (90.3% importance). This makes engineering sense because:

- Hot air is less dense, so gas turbines get less oxygen for combustion.
- Power plants produce less on hot days, which matches what we see in real life.
- The strong correlation (r = -0.948) confirms this relationship is real and consistent.

5.2 Key learnings from the project

5.2.1 Technical learning

- Different models have different strengths. No single best model for everything
- Data preparation matters. Scaling features correctly were important for SVR
- Validation is crucial. Using train-test split and cross-validation gave reliable results
- Real data has patterns. The scatter plots showed clear relationships between inputs and output

5.3 Project challenges and solutions

5.3.1 Main Challenges

- Choosing the right models.
- Understanding feature importance.
- The 452 MW case: Different models gave very different predictions for this sample.
- Making results understandable: Turning numbers into meaningful insights

5.3.2 How I Solved Them

1. I Started my models training from Linear Regression as a baseline.
2. Compared multiple approaches to see what worked best.
3. Connected ML results to engineering knowledge (thermodynamics).
4. Created clear visualizations to show patterns in the data.

6.Conclusion

In this project, I used machine learning algorithms to predict power plant output from ambient data. I tested four different models and found that Random Forest worked best overall, while Decision Tree was great for some specific cases. Also, Super Vector regression is better for low power output predictions. Temperature turned out to be the most important factor, which makes sense from what I've learned about thermodynamics. All models gave useful predictions within 1% error, showing that ML can help solve real Engineering problems. This project helped me to connect the theories of fundamental thermodynamics for Engineering applications (semester 3), power systems 1(semester 3), statistics (semester 2) and the concepts of machine learning.