Supervised Learning for Environmental Applications

From Theory to Practice in Sustainability Research

Nipun Batra
Sustainability Lab
IIT Gandhinagar

July 2025

SUSTAINABILITY LAB

Outline

What is Supervised Learning?

SUSTAINABILITY LAB

Supervised learning uses labeled training data to predict outcomes on new data.

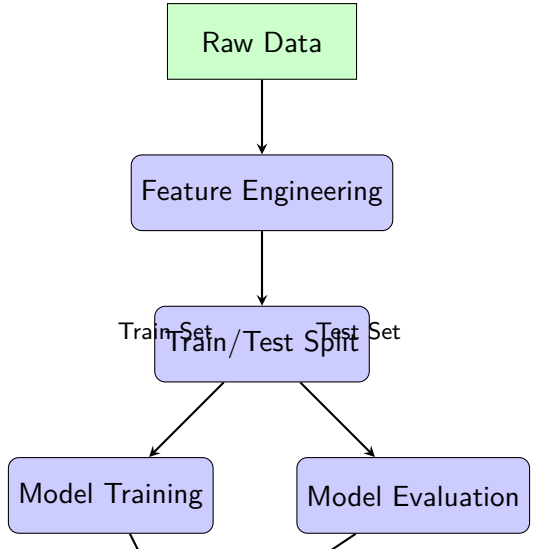
- **Classification**: Predicting discrete categories
- **Regression**: Predicting continuous values
- **Key Components**: Features, labels, model, loss function

Environmental Applications

SUSTAINABILITY LAB

- Energy consumption prediction
- Air quality classification
- Species identification from sensor data
- Climate pattern recognition

Supervised Learning Workflow



Linear Regression

Mathematical Form:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

Key Properties:

- Simple and interpretable
- Fast training and prediction
- Assumes linear relationships
- Good baseline model

Energy Consumption Example

SUSTAINABILITY LAB

Variables:

- y: Daily energy usage (kWh)
- x_1: Temperature (°C)
- x_2: Humidity (%)
- x_3: Number of occupants

Model learns how each factor influences energy consumption

Implementation Example


```python
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
import pandas as pd
from sklearn.metrics import mean_squared_error, r2_score

# Load data
data = pd.read_csv('energy_consumption.csv')
X = data[['temperature', 'humidity', 'occupancy']]
y = data['energy_kwh']

# Split and train
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)
```

Decision Trees

SUSTAINABILITY LAB

How it works:

- Splits data based on feature values
- Creates if-then rules automatically
- Handles non-linear relationships
- Easy to interpret and visualize

Advantages:

- No assumptions about data distribution
- Handles mixed data types
- Built-in feature selection

Decision Tree Example

SUSTAINABILITY LAB

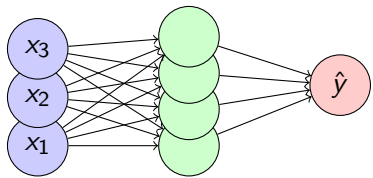
```
 if PM2.5 > 35:
  if NO2 > 40:
    class = "Poor"
  else:
    class = "Moderate"
else:
  if O3 < 80:
    class = "Good"
  else:
    class = "Moderate"
```

Neural Networks

SUSTAINABILITY LAB

Architecture:

- Interconnected layers of neurons
- Non-linear activation functions
- Learns complex patterns automatically
- Requires more data but higher accuracy



Neural Network Applications

SUSTAINABILITY LAB

Environmental Use Cases:

- **Air Quality Prediction**: Multi-pollutant forecasting
- **Energy Load Forecasting**: Grid demand prediction
- **Climate Modeling**: Temperature and precipitation patterns
- **Species Classification**: Biodiversity monitoring from audio/images

Architecture Considerations:

- **Feedforward**: Standard prediction tasks
- **LSTM/RNN**: Time series data (weather, energy)
- **CNN**: Satellite imagery analysis
- **Transformer**: Multi-modal environmental data

Performance Metrics

SUSTAINABILITY LAB

Metric	Regression	Classification
Primary	Mean Squared Error	Accuracy
Secondary	R^2 Score	F1-Score
Interpretable	Mean Absolute Error	Confusion Matrix

Key Considerations:

- Choose metrics relevant to your problem
- Consider class imbalance in classification
- Use cross-validation for robust estimates

Model Comparison

Algorithm	Energy Prediction (R^2 Score)	Air Quality (Accuracy)	Training Time (seconds)
Linear Regression	0.73	–	0.02
Logistic Regression	–	0.84	0.05
Decision Tree	0.68	0.79	0.12
Random Forest	0.81	0.87	2.45
Support Vector Machine	0.76	0.85	12.30
Neural Network	0.83	0.89	45.60

Table: 5-fold cross-validation results on sustainability datasets

Key Insights:

- Random Forest offers good balance of accuracy and speed
- Neural networks achieve highest accuracy but require more computation
- Linear models remain competitive for simpler problems

Smart Grid Optimization

SUSTAINABILITY LAB

Problem Setup:

- Predict hourly electricity demand
- Features: weather, time, historical usage
- Goal: Optimize energy generation and storage

Data Sources:

- Smart meter readings (10M+ households)
- Weather station data
- Calendar information
- Economic indicators

Grid Optimization Results

SUSTAINABILITY LAB

Model Performance:

- Accuracy: 92% within 5% error
- Cost savings: $2.3M annually
- CO2 reduction: 15,000 tons/year

Deployment:

- Real-time predictions every 15 minutes
- Automatic model retraining weekly
- Integration with grid control systems

Feature Importance Analysis

SUSTAINABILITY LAB

Feature Importance (Random Forest):

Feature	Importance
Hour of day	0.34
Temperature	0.22
Day of week	0.18
Previous day usage	0.12
Humidity	0.08
Wind speed	0.04
Holiday indicator	0.02

Key Insights

SUSTAINABILITY LAB

Analysis Results:

- Time patterns dominate (52% combined)
- Weather matters, especially temperature
- Historical usage provides context
- Complex interactions between features

Business Impact:

- Better demand forecasting
- Reduced energy waste
- Improved grid stability

Current Challenges

SUSTAINABILITY LAB

Data Quality Issues:

- Missing sensor readings
- Measurement errors and drift
- Inconsistent data collection
- Privacy and access constraints

Model Limitations:

- Assumption violations
- Overfitting to training data
- Poor generalization

Environmental Complexity

SUSTAINABILITY LAB

Environmental Challenges:

- Non-stationary patterns (climate change)
- Multi-scale temporal effects
- Spatial dependencies
- Extreme events and outliers

Practical Constraints:

- Computational resources
- Real-time requirements
- Model maintenance and updates
- Stakeholder acceptance

Future Research Directions

SUSTAINABILITY LAB

Advanced Methodologies:

- **Transfer Learning**: Adapt models across regions
- **Federated Learning**: Distributed sensor networks
- **Physics-Informed ML**: Incorporate domain knowledge
- **Uncertainty Quantification**: Better risk assessment

Integration Opportunities:

- Multi-modal data fusion
- Real-time learning and adaptation
- Human-in-the-loop decision making
- Explainable AI for policy makers

Impact Areas

SUSTAINABILITY LAB

Sustainability Applications:

- Climate change mitigation and adaptation
- Sustainable urban planning
- Renewable energy optimization
- Biodiversity conservation

Key Takeaways

SUSTAINABILITY LAB

What We've Learned:

- Supervised learning provides powerful tools for environmental problems
- Model selection depends on data characteristics and requirements
- Evaluation must consider domain-specific metrics
- Real-world deployment requires practical considerations

Best Practices:

- Start simple, then increase complexity
- Invest in data quality and understanding
- Use appropriate validation strategies
- Consider interpretability alongside accuracy

Questions and Discussion

SUSTAINABILITY LAB

How can we better leverage supervised learning for sustainability challenges?