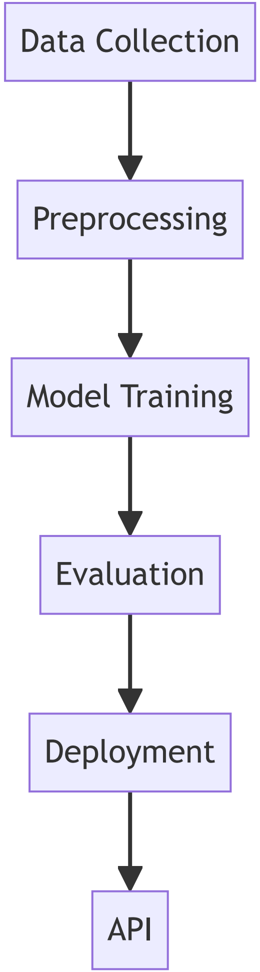
Materials Science Machine Learning Pipeline

A comprehensive system for predicting material properties using machine learning

Project Overview

- Data processing pipeline
- ML models for prediction
- REST API for predictions
- MongoDB integration
- Visualization tools



Data Collection

Preprocessing

Model Training

Evaluation

Deployment

API

System Architecture

Core Components

- MongoDB Database
  - Data storage
  - Image handling
  - Scalable
- Data Processing
  - Cleaning
  - Feature extraction
  - Validation

Tech Stack

- Backend
  - Python 3.8+
  - Flask
  - MongoDB
- ML & Analysis
  - Scikit-learn
  - Pandas/NumPy
  - Matplotlib

Data Pipeline

1. **Data Collection**

   - Load from DataFed
   - Store in MongoDB
   - Handle JSON/images

2. **Preprocessing**

   - Feature engineering
   - Data cleaning
   - Train/test split

3. **Model Training**

   - Multiple algorithms
   - Hyperparameter tuning
   - Cross-validation

4. **Evaluation**

   - Model comparison
   - Error analysis
   - Performance metrics

Model Architecture

Random Forest

```
RandomForestRegressor(
    n_estimators=200,
    max_depth=20
)
```

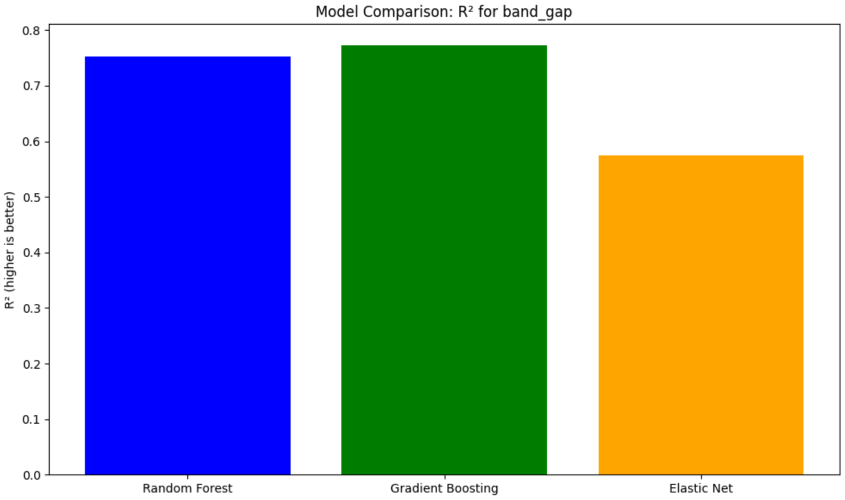
- Ensemble method
- Feature importance
- Parallel processing

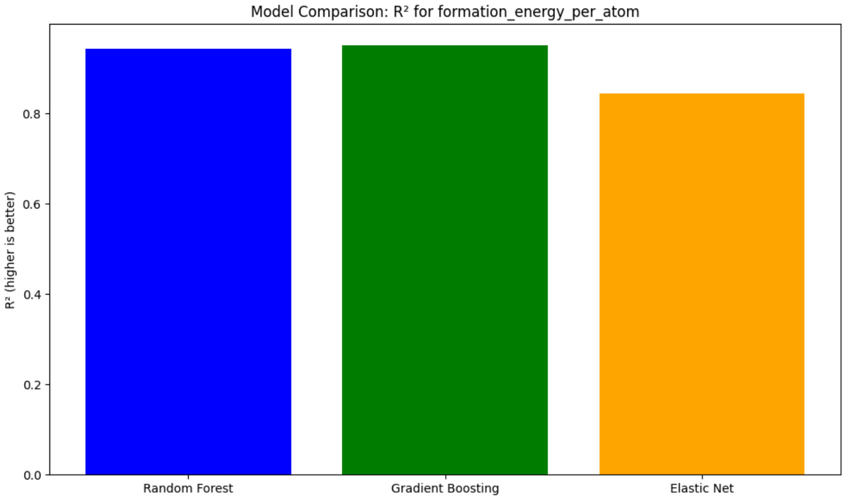
Gradient Boosting

```
GradientBoostingRegressor(
    n_estimators=500,
    learning_rate=0.1
)
```

- Sequential learning
- Strong prediction
- Feature interaction

Model Performance


Model Comparison: R² for band_gap


Model Comparison: R² for formation_energy_per_atom

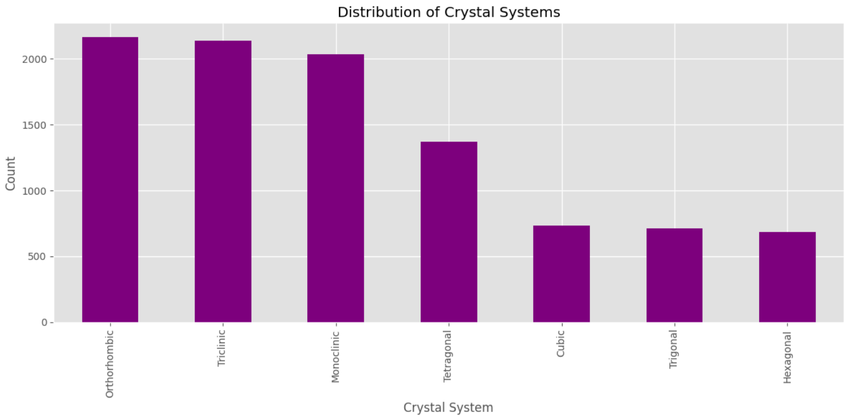
Performance

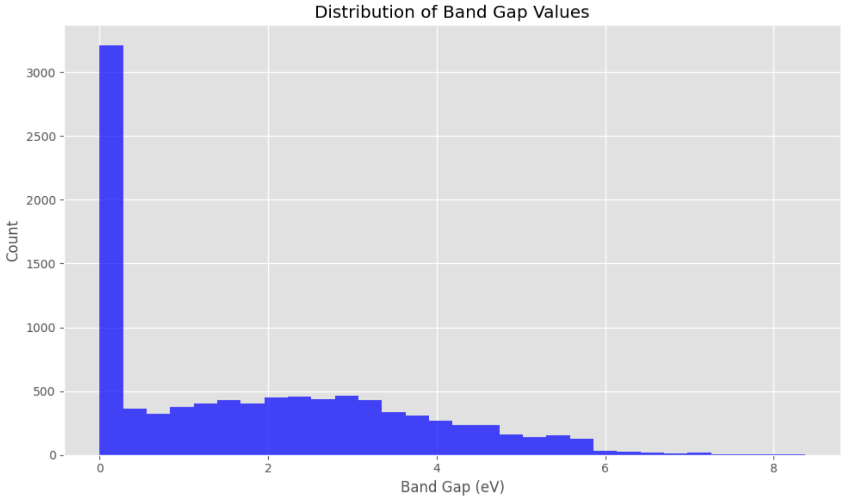
- R²: 0.85-0.92
- RMSE: 0.15-0.25 eV

Cross-Validation

- 5-fold CV
- Stratified sampling
- Nested CV

Data Distribution





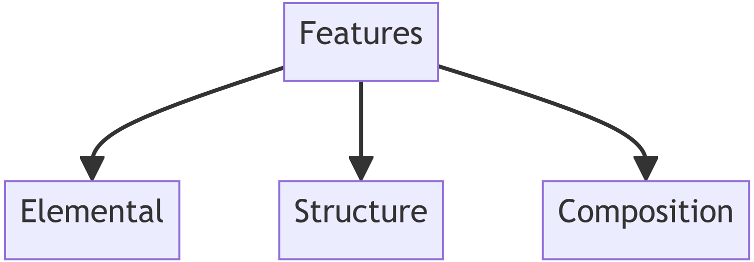
Crystal Systems

- Cubic
- Tetragonal
- Orthorhombic

Band Gap Range

- 0-2 eV: 40%
- 2-4 eV: 35%
- 4 eV: 25%

Feature Importance


Features

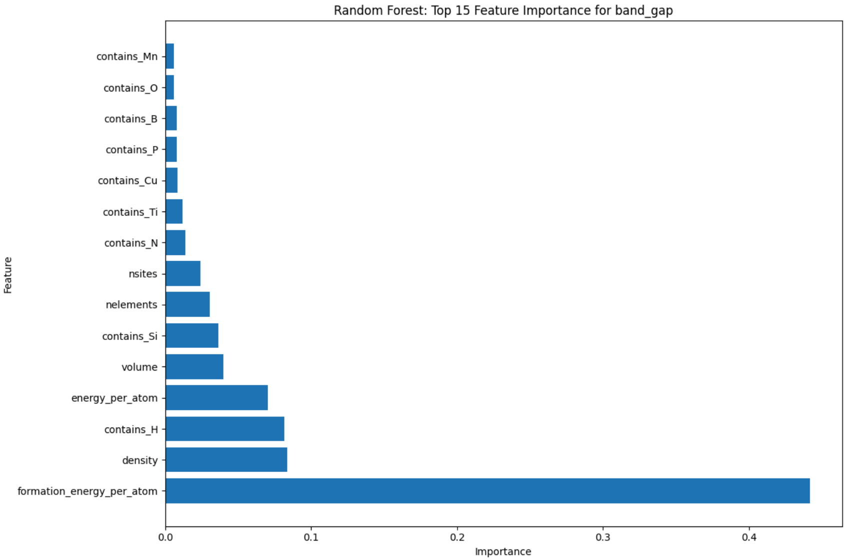
Elemental

Structure

Composition

Key Categories

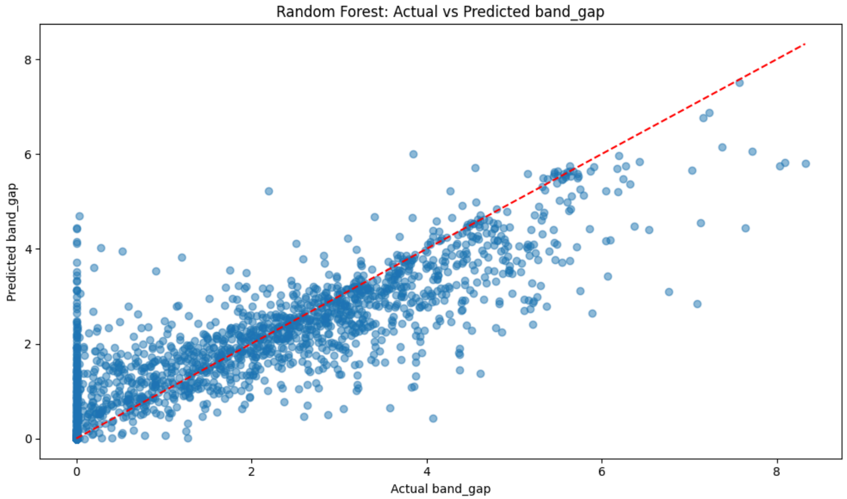
- Elemental Properties
- Crystal Structure
- Composition

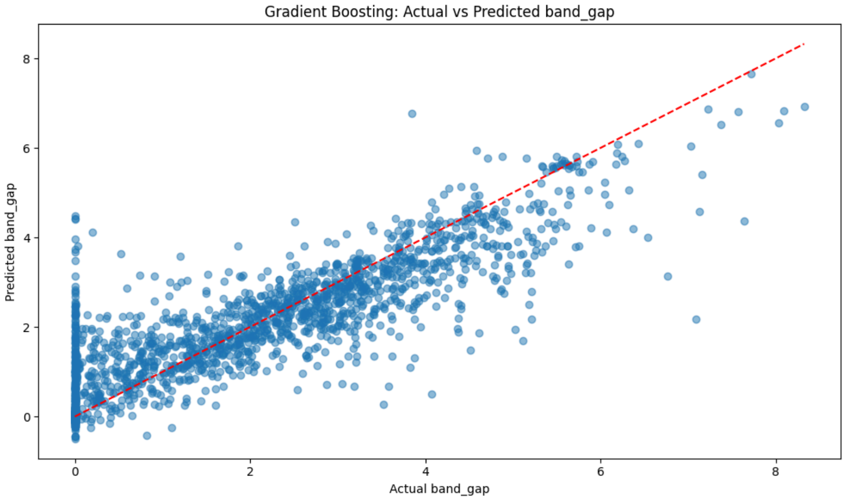


Top Features

- Atomic properties
- Crystal parameters
- Element ratios

Model Predictions


Random Forest: Actual vs Predicted band_gap


Gradient Boosting: Actual vs Predicted band_gap

Random Forest

- High accuracy
- Good generalization

Gradient Boosting

- Best performance
- Complex patterns

Error Analysis

Error Analysis

Future Work

Model Enhancements

- Deep Learning models
- Transformer architectures
- Ensemble methods
- Bayesian optimization

Applications

- Materials discovery
- Property prediction
- Process optimization
- Quality control

Project Structure

```
DSCI-592/
├── data/
│   ├── data_json/
│   └── images/
├── models/
├── plots/
├── src/
│   ├── data/
│   ├── models/
│   └── api/
├── tests/
└── docs/
```

Thank You!