National Textile University, Faisalabad

Department of Computer Science

Name:	Muhammad Bilal Rafique
Registration No:	22-NTU-CS-1192
Class:	BSCS-7B
Course Name:	Machine Learning
Lab Plan	#07
Submitted To:	Miss Kainat Abdullah
Submission Date:	Dec 7, 2025

LAB REPORT 07

Understanding Ensemble Models

1. Project Introduction & Goal :-

The primary objective of this machine learning project is to explore **ensemble learning techniques** and evaluate their effectiveness in predicting electricity price changes in the New South Wales electricity market. Ensemble models combine multiple base models to improve predictive performance and reduce errors that may arise from relying on a single model.

Specifically, this project focuses on a **classification task**: predicting whether the electricity price will go **UP** or **DOWN** in New South Wales, based on a moving average of the last 24 hours. The dataset spans 45,312 instances collected from 7 May 1996 to 5 December 1998, with each record representing a 30-minute interval. Features include time, day, electricity demand in New South Wales and Victoria, price, and scheduled electricity transfer between states.

2. Explaining Ensemble Learning:-

An **ensemble model** is a machine learning approach where multiple models are combined to make a prediction. The idea is that aggregating the outputs of different models often yields more accurate and robust predictions than using any single model alone. This is sometimes referred to as the "wisdom of the crowd" analogy: a collective decision from multiple predictors can outperform an individual expert.

Ensemble methods are broadly categorized into three types:

1. **Averaging Methods** – These methods train multiple base models independently and combine their predictions by averaging (for regression) or voting (for classification). Examples include **Bagging**, **Random Forests**, and **Extra Trees**. Bagging reduces variance by training models on different subsets of data.
2. **Boosting Methods** – Boosting trains models sequentially, with each new model focusing on correcting the errors of previous ones. It aims to reduce bias and improve performance on difficult examples. Examples include **Gradient Boosting**, **AdaBoost**, and **XGBoost**.

3. **Stacking** – Instead of averaging predictions, stacking trains multiple models on the full dataset and then trains a meta-model on their outputs. The meta-model learns how to combine the base model predictions optimally.

This project uses ensemble models in the **classification domain** to highlight their benefits and performance.

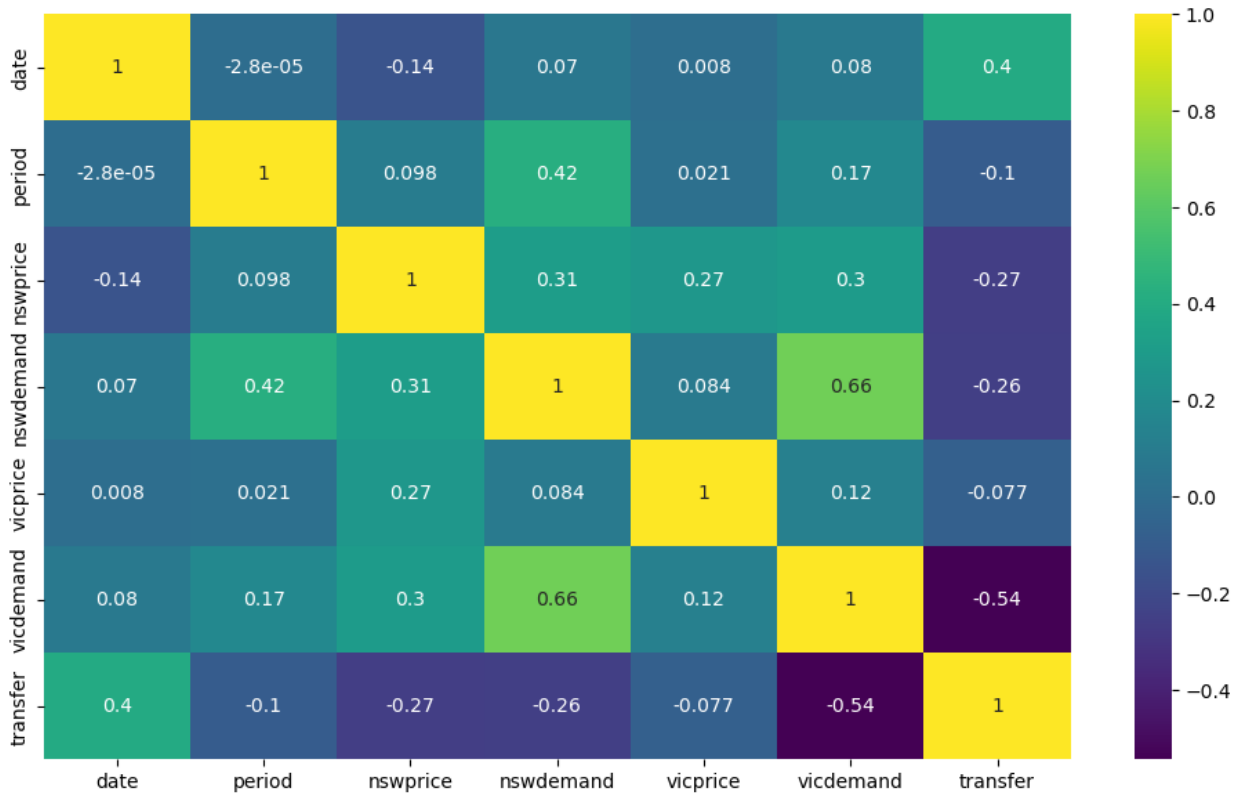
3. Data Exploration & Insights:-

The dataset used comes from the **Australian New South Wales Electricity Market**, available on OpenML. It contains features such as date, day, period, NSWprice, NSWdemand, VICprice, VICdemand, and transfer. The **target variable** is class, which indicates whether the electricity price is going **UP** or **DOWN** relative to a 24-hour moving average.

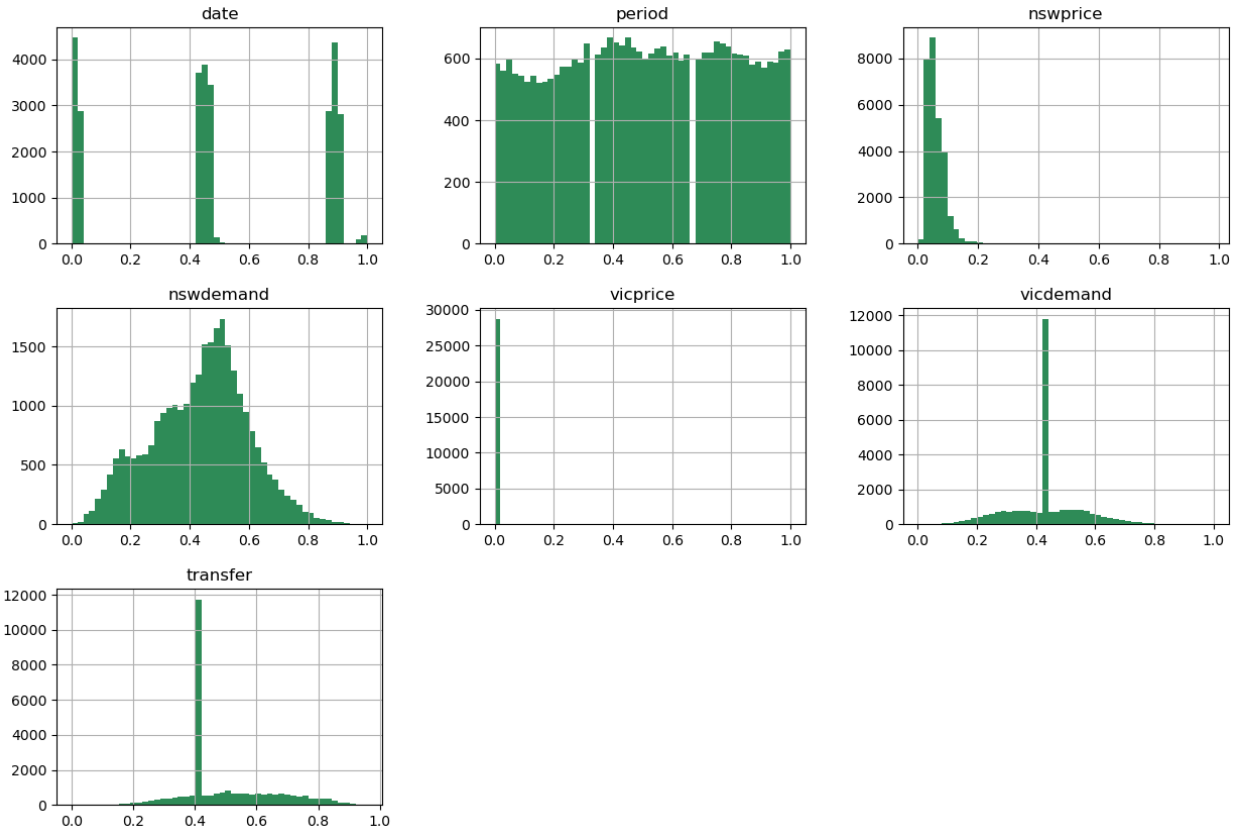
EDA revealed several insights:

1. The **class distribution** is relatively balanced, with both UP and DOWN occurrences well represented in the training set. This ensures that models can learn patterns for both outcomes effectively.
2. The correlation analysis indicated that features such as **NSW demand** and **NSW price** have moderate relationships with each other, while electricity transfer shows some dependency on both demand and price. Visualizations also showed day-wise variations in price trends.

These insights help in understanding which features might be more influential for predicting electricity price changes.



Each Feature is correlated with itself. If one feature is closer to +1 it means they both have similar information Like vicedeman and nswdemand has 0.66. And if one feature is closer to -1 it means they share different information.



4. Methodology & Planned Workflow:-

The notebook implements several ensemble algorithms, including **Bagging Classifier**, **Gradient Boosting Classifier**, and **AdaBoost Classifier**.

A typical workflow for training these models in Scikit-learn involves the following steps:

1. **Data Preprocessing** – Clean the data, encode categorical variables (e.g., label encoding for the target), and split into training and testing sets.
2. **Model Initialization** – Define the ensemble model with appropriate hyperparameters, such as number of estimators, learning rate, and maximum depth.
3. **Training** – Fit the model on the training dataset using .fit().
4. **Prediction & Evaluation** – Use the trained model to predict on training and test data, then evaluate performance using metrics like **accuracy** and **cross-validation scores**.
5. **Hyperparameter Tuning** – Optionally, adjust parameters to optimize model performance.

For instance, the **Gradient Boosting Classifier** in this notebook was trained with 450 estimators, a learning rate of 0.5, and a max depth of 2. Cross-validation showed a mean accuracy of approximately 88% on training data.

5. Conclusion & Reasoning:-

Ensemble models are particularly suitable for predicting electricity price changes because they combine multiple learners to reduce errors and handle the complex, fluctuating patterns of electricity markets. Unlike single models, ensembles can capture both linear and nonlinear relationships more effectively, leading to improved generalization on unseen data.

To assess the success of the models, the **most important evaluation metric** is the accuracy on the test set. Additionally, cross-validation scores and comparison across different ensemble methods help identify the most robust model. In this project, the **Gradient Boosting Classifier** achieved the highest test accuracy (~88%), indicating its effectiveness for this classification problem