Name:	Syed Junaid Jaffery
Class:	BSCS-A-7
Registration No:	22-NTU-CS-1167
Lab Report:	Ensemble Learning Report
Course Code:	AIC-3072L
Course Name:	Machine Learning Lab
Submitted To:	*Ms. Kainat Abdullah*
Submission Date:	December 14, 2025

1. Project Introduction and Goals

What is the main objectve of this machine learning project?

Its basically an intro to ensemble models with an example of binary classification being done using several ensemble models. The main objective is to gain an understanding of different types of ensemble methods and which could be used when.

What is the specific task the model is being built to perform?

He is trying to do binary classification on a dataset, The specific task is to correctly predict the class label which is either UP or DOWN, as in electricity prices going up or down.

2. Explaining Ensemble Learning

In simple terms, what is an "ensemble model"?

An ensemble is simply a combination of multiple models to end up with a stronger overall model. The voting classifier that we have studied in Labs is an example of this actually so we have covered this to an extent already.

The notebook mentions three main categories of ensemble methods (Averaging, Boosting, Stacking). Briefly describe what each category does.

1. Averaging (also known as bagging) is the simplest, we train multiple models of the same type (almost always decision trees) on random subsets of the dataset (for randomization, sampling with replacement technique is used), completly independent from eachother. We make predictions using each model, the majority vote wins.
2. Boosting is when the models are not training independently, one model is supposed to specifically target the samples where the first model had high error. If we have two models x and y and we wanna train them using the boosting technique, model y would specifically train on the samples that model x failed on so they both combine to form a more robust model. Each model builds on the previous.
3. Stacking is a 2 step process, first is the training of the base models, and the second step is a manager model trained on the predictions of the base models. The base models are diverse (in bagging and boosting we just use a bunch of models of the

same type, mostly decision trees, but in stacking we gotta use diff types of models like svm + knn + decision tree + a simple ANN) and they provide their own independent predictions. A manager model is then trained on those predictions to provide the final prediction.

---

3. Data Exploration & Insights

Describe the dataset used. What is the source and what does the target variable represent?

- Description: its a dataset about electricity with 45,312 samples and 8 features and one target class so a total of 9 columns. All the features are already numeric, the target class is categorical.
- Source: its electrity dataset version 1 from the OpenMl website. You can import it directly from sklearn using sklearn.datasets.fetch_openml(name="electricity", version=1)
- Target: its the binary class label of either UP or DOWN indicating whether the electricity price has increased or decreased

Based on the Exploratory Data Analysis (EDA), what are two key findings about the data? (For example, discuss the distribution of the target variable or a relationship between features).

- There are more DOWN label samples then UP label but it is not too heavily skewed that we need something like stratefied cross validation.
- None of the features have such a high correlation that they become basically redundant so all features are valid. (although day is not normalized)

---

4. Methodology & Planned Workflow

The notebook lists several specific ensemble algorithms to be implemented (e.g., Gradient Boosting, AdaBoost). Name two of them.

You just named one in the question (adaboost) so thats one. I will list a few below:

1. Bagging: RandomForest and BaggingClassifier from sklearn.ensemble.BaggingClassifier with KNN as the estimator.

2. Boosting: AdaBoost, XGboost
3. Stacking: stacking classifier from sklearn.ensemble with svc, knn, and random forest as base models, and logistic regressor as manager

The code for training the models is not shown. Based on a typical Scikit-learn workflow, what are the main steps you would take to train one of these models after the data is preprocessed?

Here are the main steps after data preprocessing:

1. split the data
2. pick an ensemble model (eg random forest)
3. set hyperparameters like n_estimators and learning rate etc.
4. fit it on the training set
5. predict on the testing set
6. use accuracy_score from sklearn.metrics for quick evaluation

---

5. Conclusion & Reasoning

Why do you think ensemble models are a suitable choice for predicting electricity price changes?

The main issue that jumps to mind is that the data is complex and we need complex models to give reasonable accuracy, but complex models lead to overfitting in most cases. That scenario is perfect for using bagging so that is why ensemble methods (specifically bagging) would be perfect here. If the data is too complex for a bagging ensemble then boosting could be used since it does better then bagging on complex data. Whether you use bagging or boosting, both would generalize better then a single model.

If you were to complete this project, what would be the most important thing to check to see if the model is successful?

I don't know what exactly you mean by 'complete this project' but the logical next steps from where the notebook ends would be:

- He checked accuracy but didn't graph a learning curve to check for overfitting. He didn't even use anyother metric other then accuracy. We should at least visualize a confusion matrix.
- No hyperparameter tuning was actually done, you could use a validation set and try out different values of n_estimators etc to find out the best one (or just do gridsearchCV).

Word Count of the answers totaled: **731**

its only 31 words over the given limit, hopefully not an issue.