Classification Modeling and Evaluation

Group 6

22000415 Yang Chan
22000689 Jeong Yisak
22100809 Hwang Eunji
22100173 Kim Jaehee

HANDONG GLOBAL UNIVERSITY

Contents

HANDONG GLOBAL UNIVERSITY

Reference : 상수관로 누수 감지 데이터, ㈜유솔, AI-HUB (2020)
:https://www.andonginews.com/news/articleView.html?idxno=16895

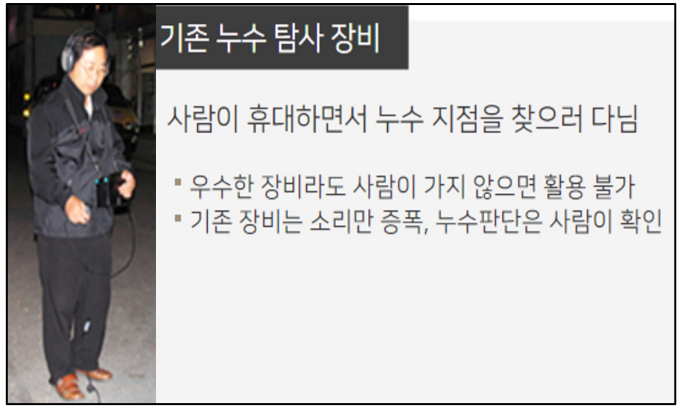
➢ Problem definition

5년간 상수도 누수로 3조3천억 원 새어나가

권기일 기자 | 승인 2023.08.17 09:10

대도시 비해 지방은 매우 높아 지역별 편차 매우 커
김 의원, 누수율 줄이고, 지역 편차 줄일 방안 강구할 것

5년간 전국 상수도에서 올림픽 규격 수영장(2,500㎥) 139만 3천개를 가득 채울 정도의 깨끗한 물이 새어 나간 것으로 나타났다. 금액으로 환산하면 3조 3천억원에 달한다.

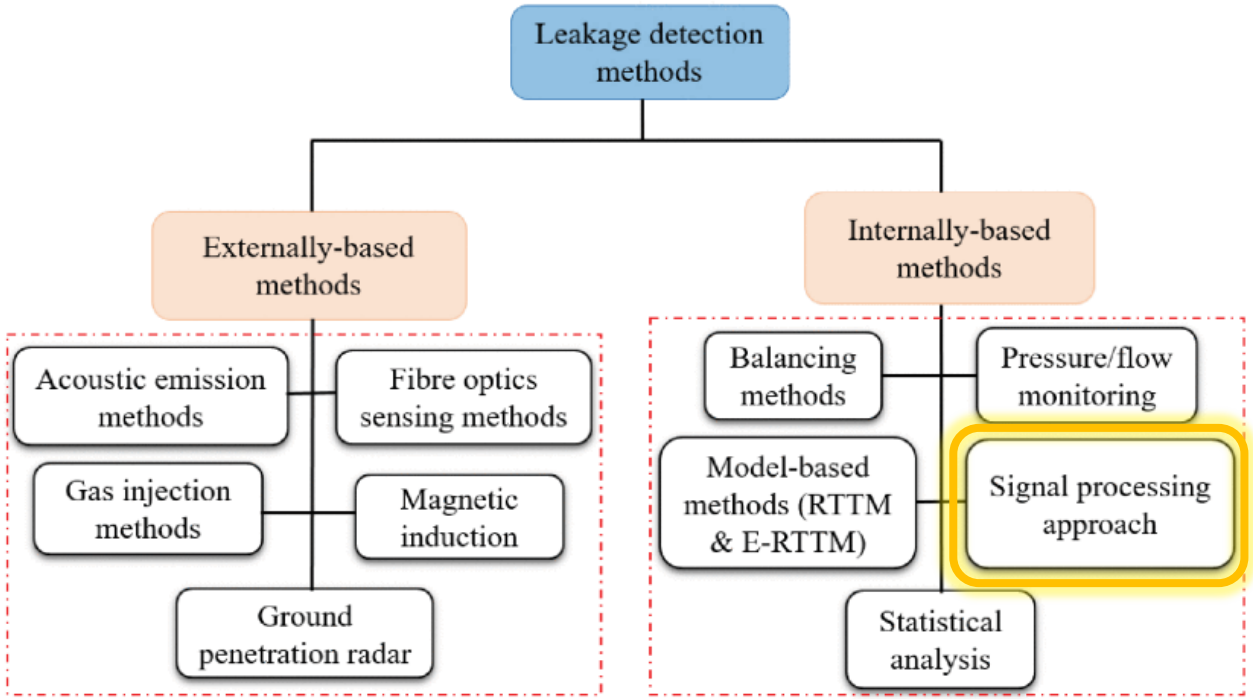


The existing method for detecting leaks required the experts at the leak site.

Various costs and significant annual losses occurred due to leak issues.

Research on leak detection systems using sensor data is needed to reduce detection time and costs.

➢ Data Processing Procedure



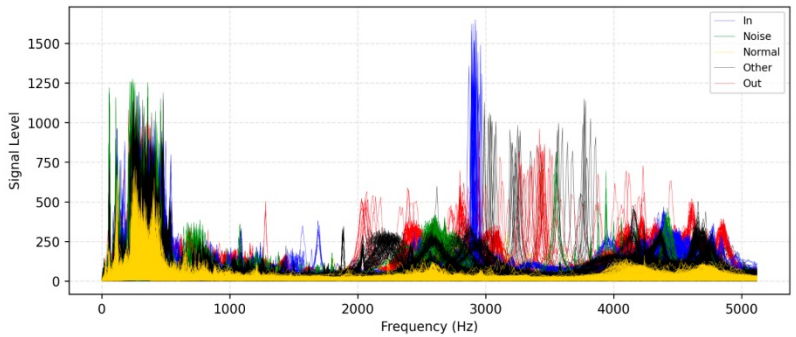
Kazeem B. Adedeji et al., *IEEE* (2017)

Reference : 상수관로 누수 감지 데이터, ㈜유솔, AI-HUB (2020)

Raw Data Collection

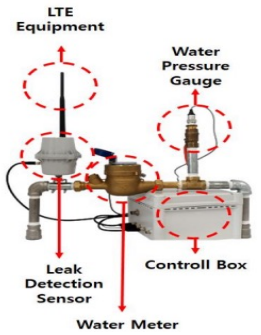
Leak Detection through On-site Investigation - Labeling

Leak Detection Using Signal Processed Data



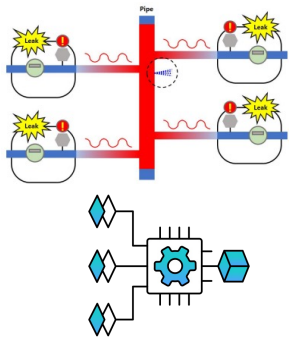
01 | Introduction

➢ Data Collection Process



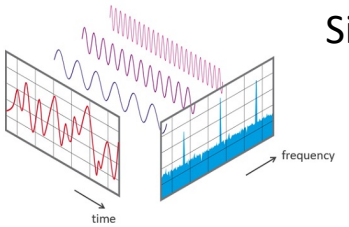
Flow and Water Pressure Monitoring System

- The sensors are installed in drainage pipes and the data is stored on the server using battery and LTE communication
- Leak detection sensors installed in a total of 11,000 locations at Gwangju and Jeollanam-do.



Data Cleaning and Labeling Process by Class

- Data cleaning is performed based on the results confirmed on-site through actual leak investigations.
- Labeling is applied to each data point according to its class.



Signal Processing Procedure

- Extract frequency spectrum density data by applying the Fast Fourier Transform (FFT) to the sound data.
- Build the dataset with the processed frequency spectrum density data.

* NTI-audio

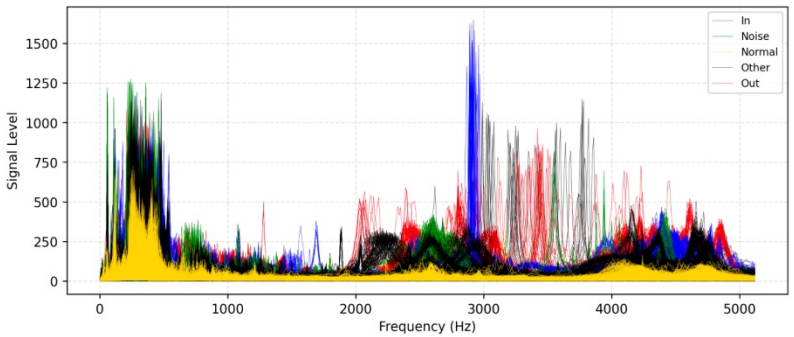
Jungyu Choi et al., *The Institute of Electronics and Information Engineers* (2023)

Reference : 상수관로 누수 감지 데이터, ㈜유솔, AI-HUB (2020)

Raw Data Collection

Leak Detection through On-site Investigation - Labeling

Leak Detection Using Signal Processed Data

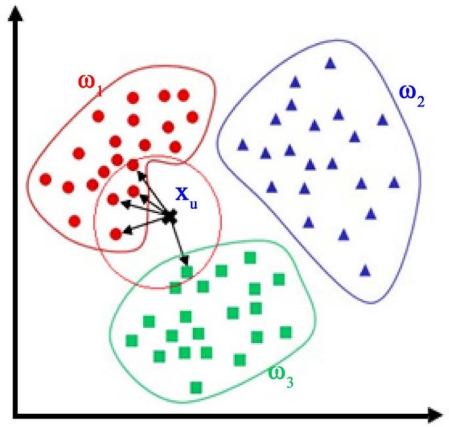


HANDONG GLOBAL UNIVERSITY

➢ Classification Algorithms

K-NN

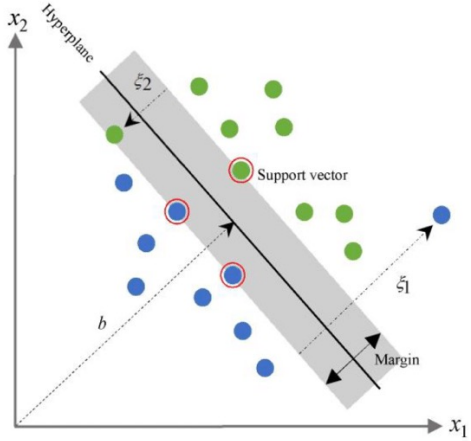
Classify by finding the K nearest neighbors based on the distance



- Suitable for **small datasets**
- **Performance drops with many variables**

SVM

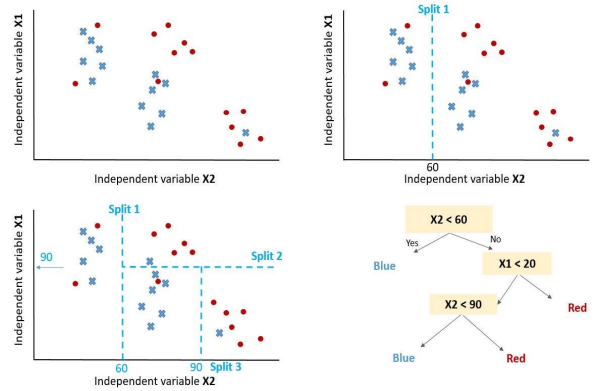
Classify by finding the hyperplane that best separates the data



- Suitable for **high-dimension data**
- **Requires parameter tuning** and has **high computational cost**

Decision Tree

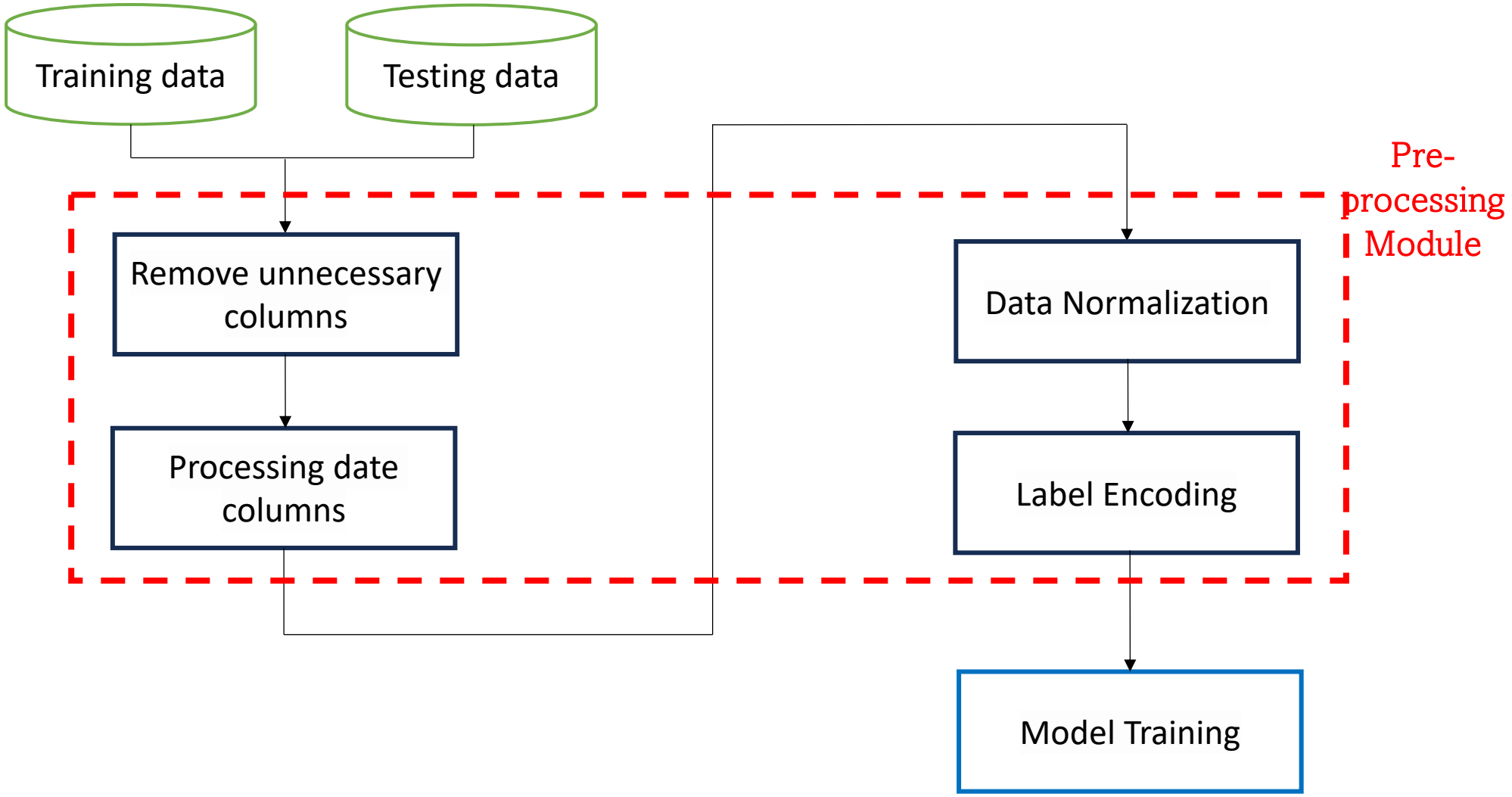
Classify using a tree structure that makes decisions



- **Easy to understand** the features and outcomes of the data
- Can be **prone to overfitting**

03 | Algorithm Description

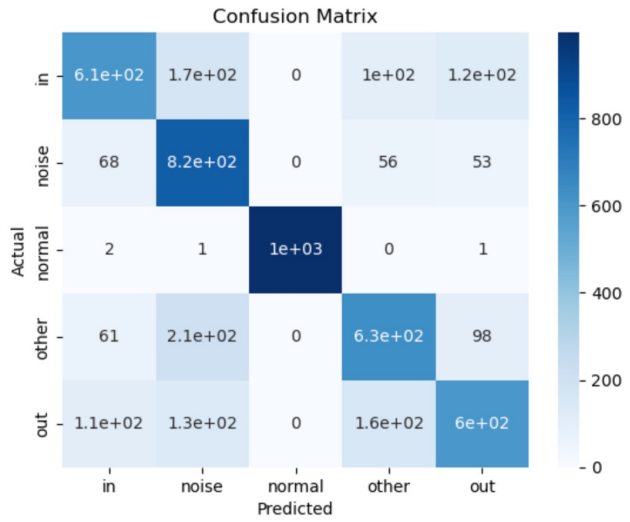
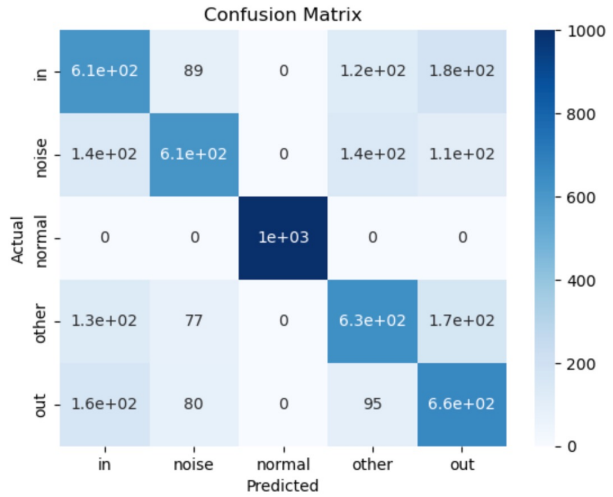
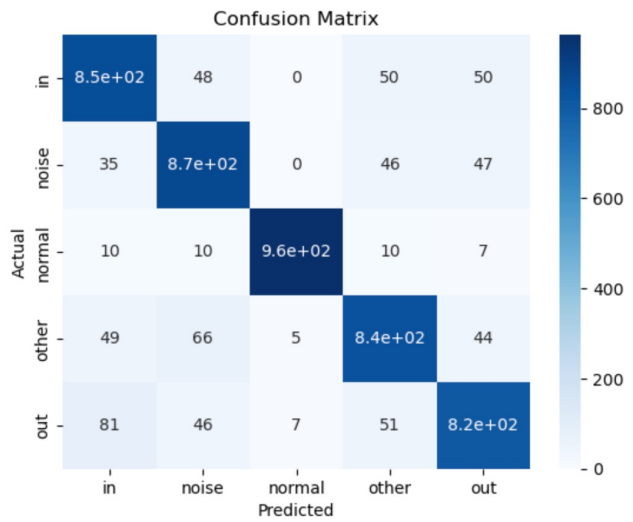
➢ Data Pre-processing



03 | Algorithm Description

➢ Processing for Selecting the algorithm

① Accuracy values comparison

SVM	Decision Tree	K-NN
		
Accuracy: 0.7300	**Accuracy: 0.7030**	**Accuracy: 0.8676**

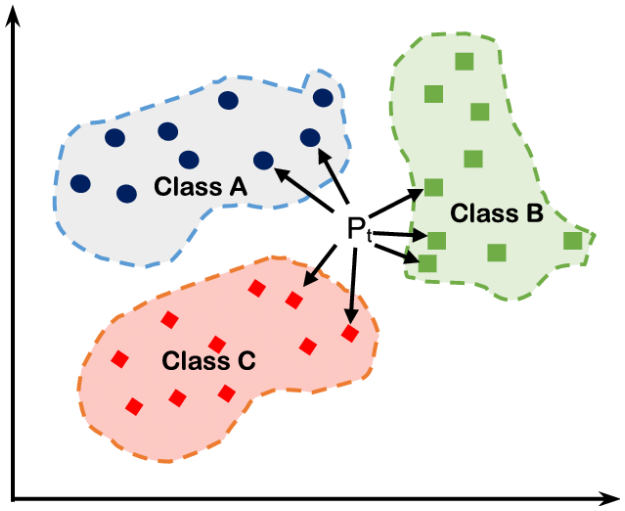
➢ Processing for Selecting the algorithm

②−1 Analysis of the K-NN Algorithm

K-NN Algorithm makes class differentiation easy.

If the classes are clearly differentiated based on leak type, K-NN can easily learn these patterns and effectively make predictions for each data point.

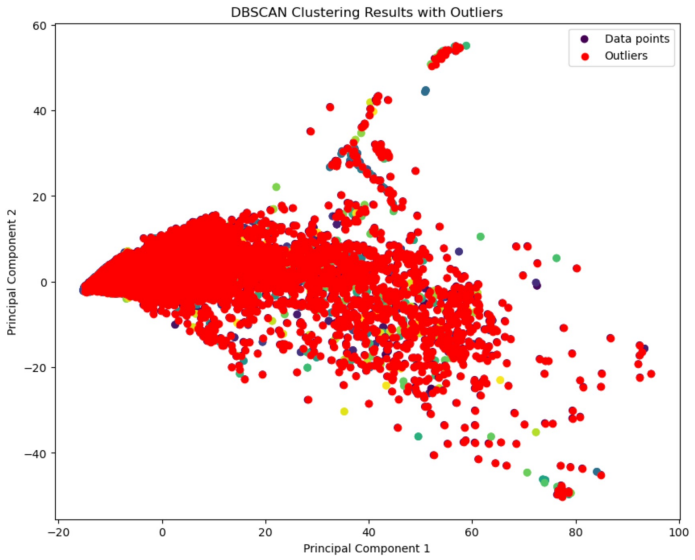
The clearer the boundaries between classes, the better the performance of K-NN.



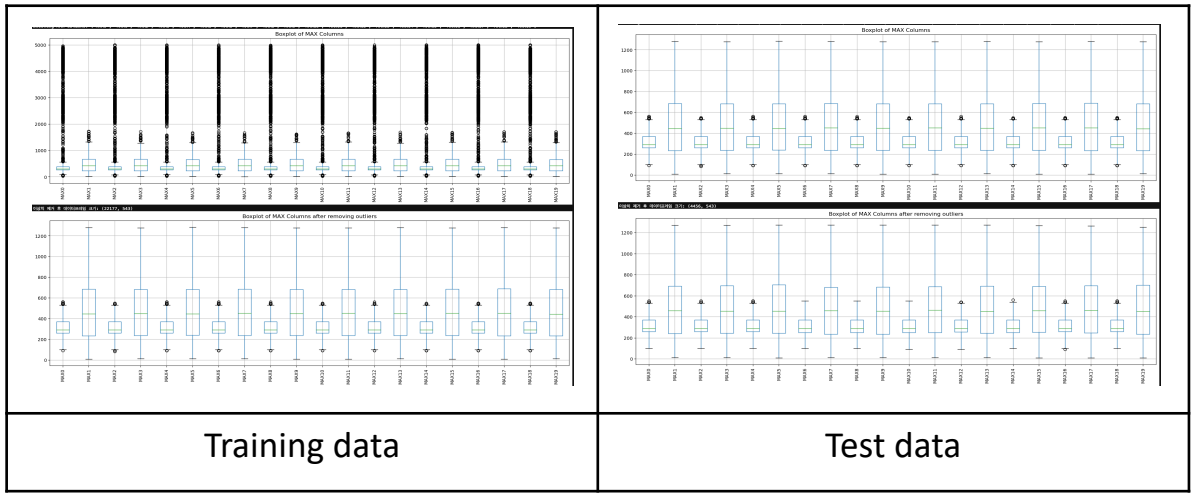
➢ Processing for Selecting the algorithm

②−2 Analysis of the K-NN Algorithm

K-NN is sensitive to outliers, so it performs better when there are fewer outliers.



- Extract the numerical to check the distribution of outliers (using DBSCAN)
- Numerous outliers are observed.

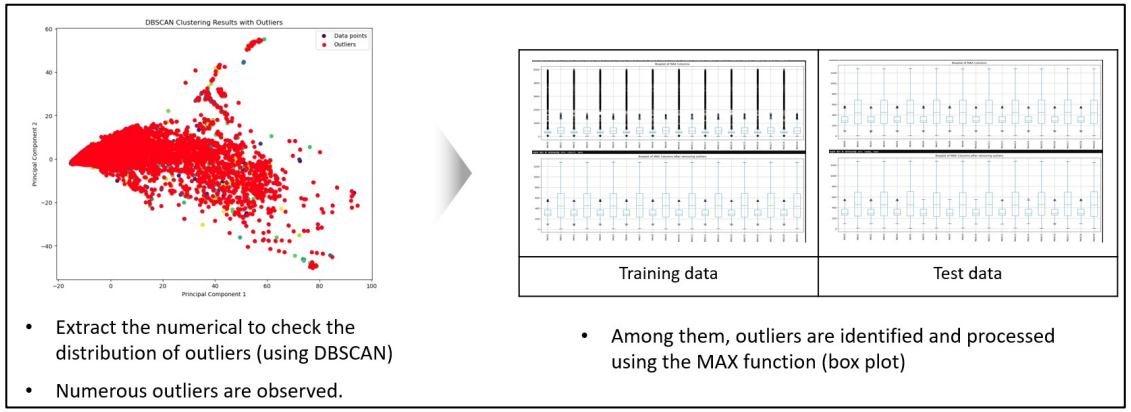


- Among them, outliers are identified and processed using the MAX function (box plot)

➢ Processing for Selecting the algorithm

②−2 Analysis of the K-NN Algorithm

K-NN is sensitive to outliers, so it performs better when there are fewer outliers.



- Extract the numerical to check the distribution of outliers (using DBSCAN)
- Numerous outliers are observed.

- Among them, outliers are identified and processed using the MAX function (box plot)

Before outlier removal	After outlier removal
Accuracy: **0.8678**	Accuracy: **0.8200**

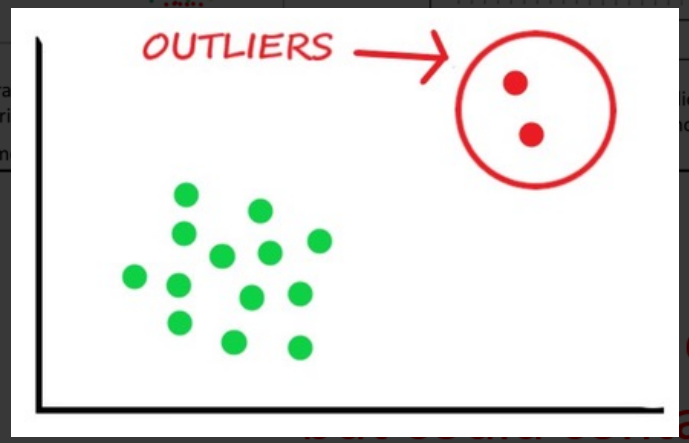
Outliers in the dataset may not simply be incorrect data points, but could contain important information about the issue at hand.

➢ Outlier

➢ Processing for Selecting the algorithm

②−2 Analysis of the K-NN Algorithm

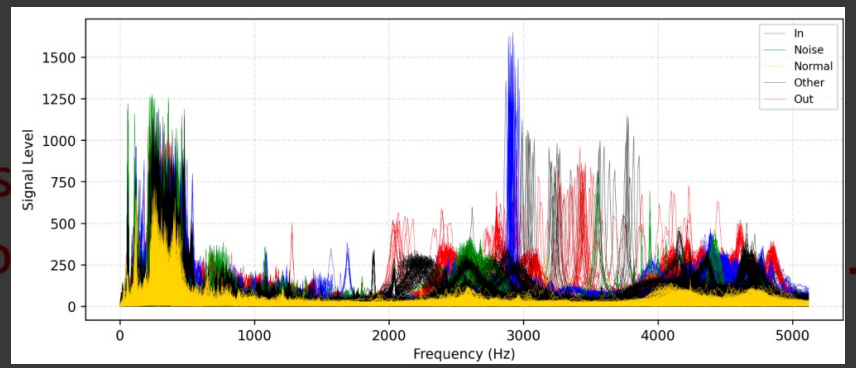
Values that deviate from the common patterns in the data, which can sometimes occur due to measurement errors or recording mistakes.



➢ Important Signals

Data points that reflect actual phenomena or conditions requiring special attention in signal processing; such as 'out', 'in', and ‘other' in our dataset.

These classifications indicate that the data reflects specific situations, and this information can be important in real-world applications.

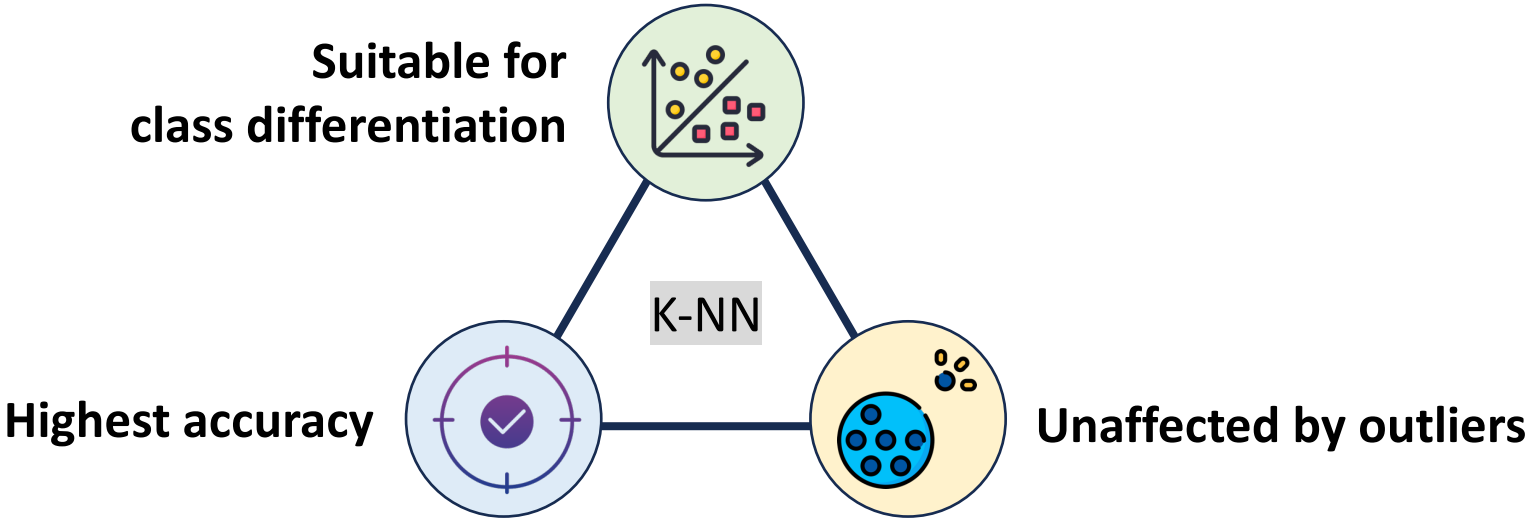


03 | Algorithm Description

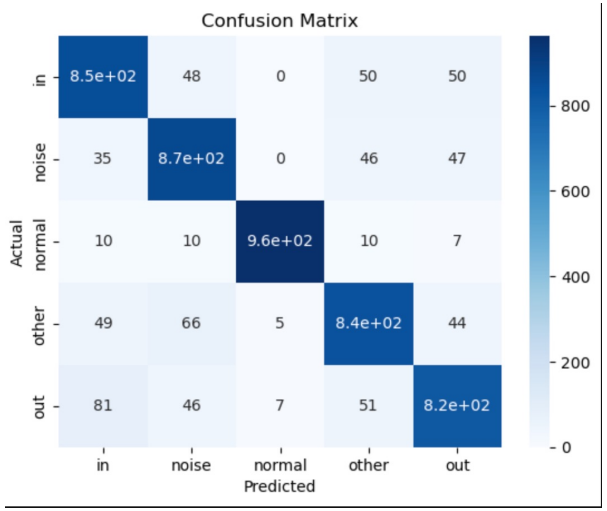
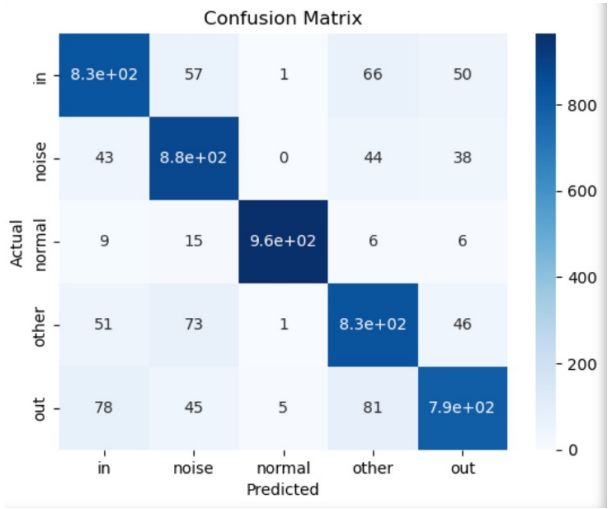
➢ Processing for Selecting the algorithm

✓ We chose K-NN Algorithm as the classification model.

This data is easily classified by leak type, and although there are outliers, it is appropriate to consider them as important signals based on the leak type rather than as mere outliers.



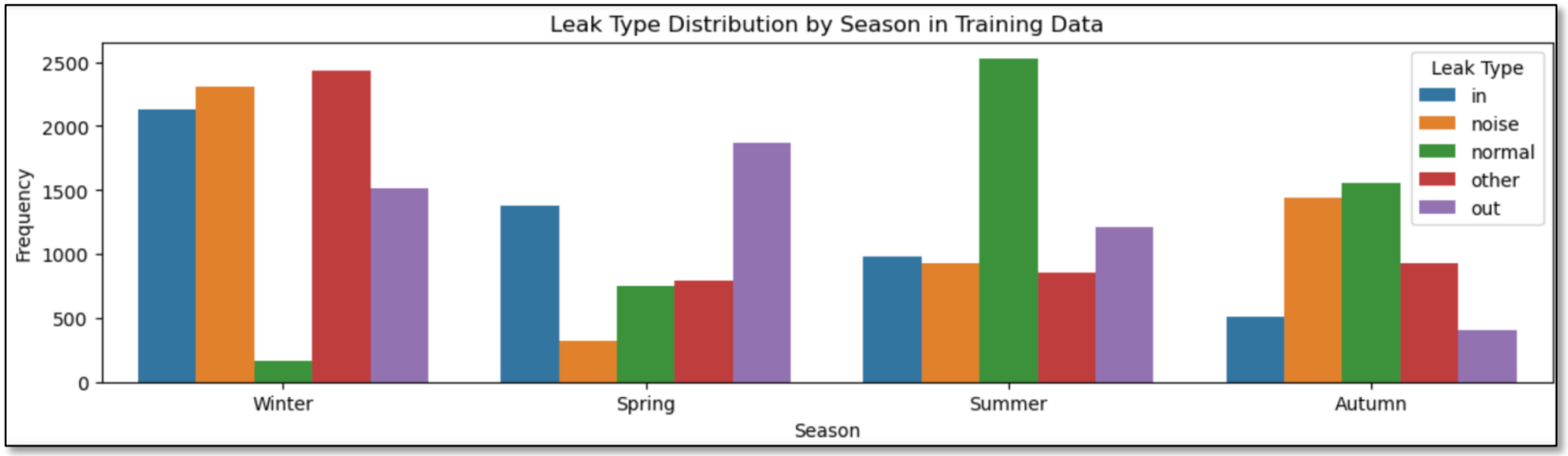
① Set K value

K=3	K=5
	
Accuracy: 0.8676 [in] : Precision: 0.83, Recall: 0.85 [noise] : Precision: 0.84, Recall: 0.87 [normal] : Precision: 0.99, Recall: 0.96 [other] : Precision: 0.84, Recall: 0.84 [out] : Precision: 0.85, Recall: 0.81	**Accuracy: 0.8570** [in] : Precision: 0.82, Recall: 0.83 [noise] : Precision: 0.82, Recall: 0.88 [normal] : Precision: 0.99, Recall: 0.96 [other] : Precision: 0.81, Recall: 0.83 [out] : Precision: 0.85, Recall: 0.79

② Categorize the ldate variable by season

Categorize the `ldate` variable by season to create new `season` variables

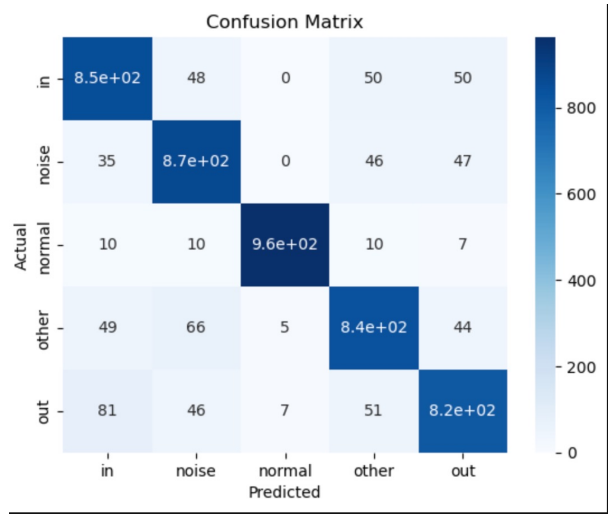
✓ We thought that leaks in the pipelines might be related to the weather.



The distribution of each leak type varies by season in the dataset.

② Categorize the Idate variable by season

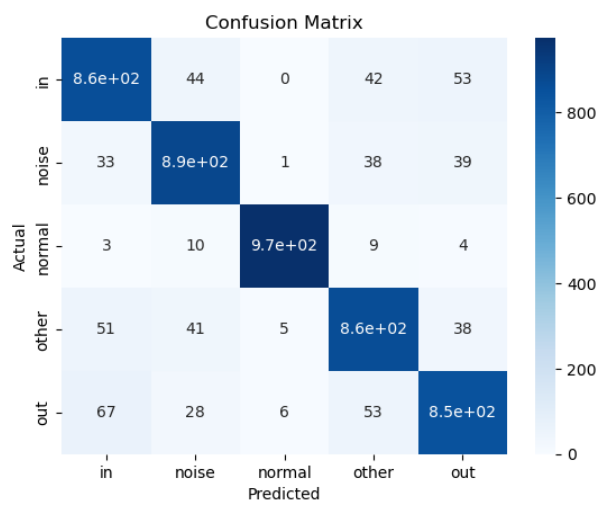
With Original data



Accuracy: 0.8676

[in] : Precision: 0.83, Recall: 0.85
[noise] : Precision: 0.84, Recall: 0.87
[normal] : Precision: 0.99, Recall: 0.96
[other] : Precision: 0.84, Recall: 0.84
[out] : Precision: 0.85, Recall: 0.81

Add Season Column



Accuracy: 0.8870

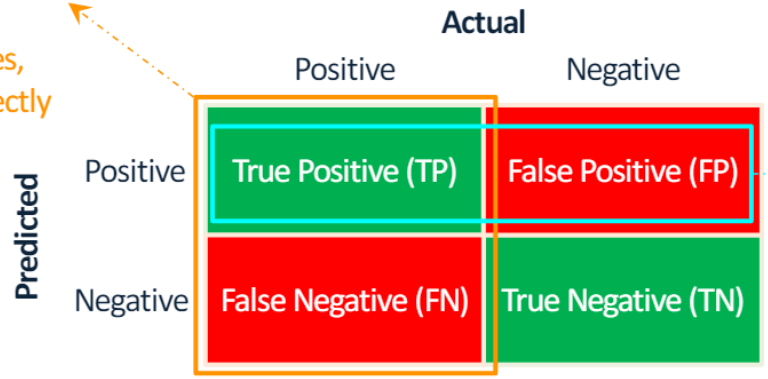
[in] : Precision: 0.85, Recall: 0.86
[noise] : Precision: 0.88, Recall: 0.89
[normal] : Precision: 0.99, Recall: 0.97
[other] : Precision: 0.86, Recall: 0.86
[out] : Precision: 0.86, Recall: 0.85

➢ How can we evaluate a classification model

- **Accuracy** shows how often a classification ML model is correct **overall**.
- **Precision** shows how often an ML model is correct when **predicting the target class.**
- **Recall** shows whether an ML model can find **all objects of the target class**.

$$Recall = \frac{TP}{TP + FN}$$

From all the actual positive classes, how many the model predicts correctly



$$Precision = \frac{TP}{TP + FP}$$

From all the predicted positive classes, how many are actually positive

$$Accuracy = \frac{TP + TN}{TOTAL}$$

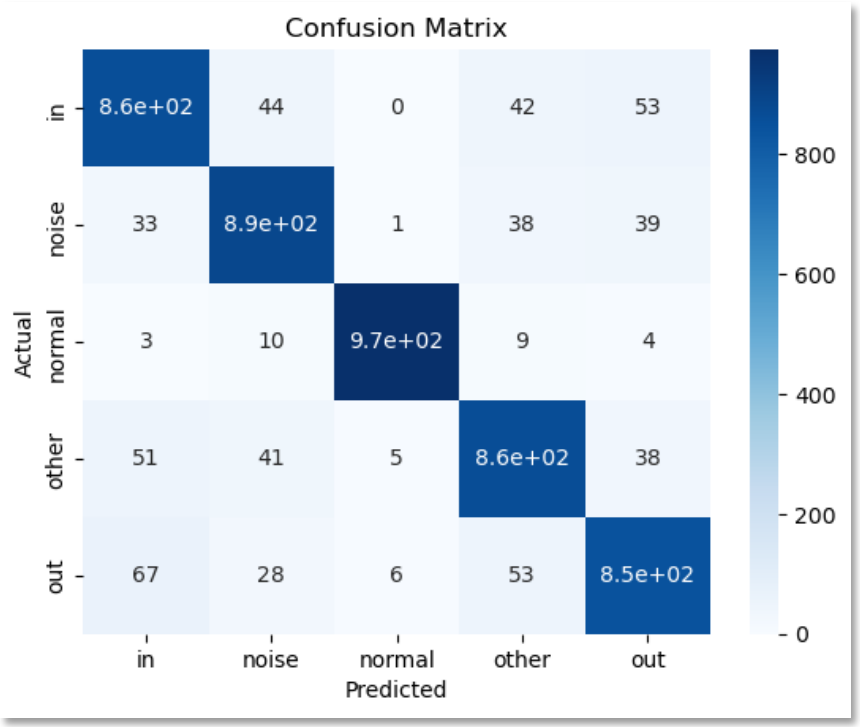
From all the classes (positive and negative), how many of them we have predicted correctly

HANDONG GLOBAL UNIVERSITY

05 | Evaluation

- ✓ We made Fusion Matrix by comparing the model test results with the labeled data.

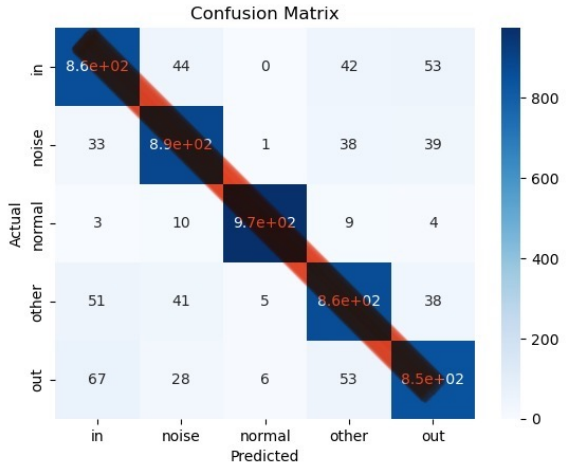
```
Confusion Matrix:
[[861  44   0  42  53]
 [ 33 889   1  38  39]
 [  3  10 974   9   4]
 [ 51  41   5 865  38]
 [ 67  28   6  53 846]]
```



05 | Evaluation

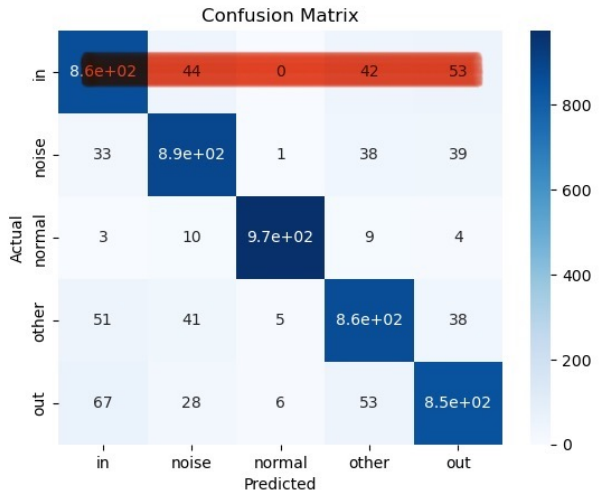
✓ We created a function that calculates accuracy, precision, recall and represented each leak type.

Accuracy



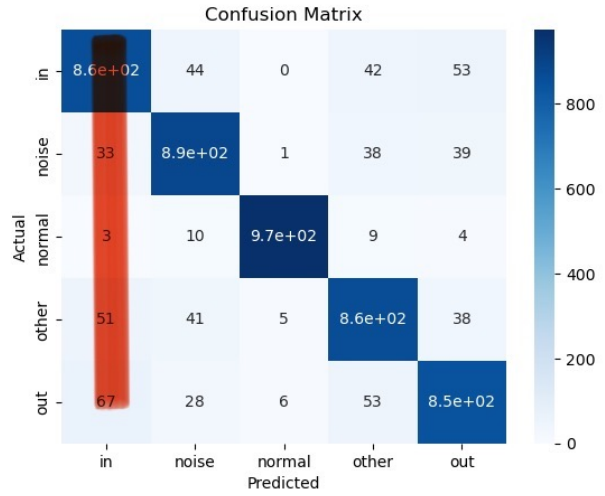
$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{All predictions}}$$

Precision



$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives + False Positives}}$$

Recall

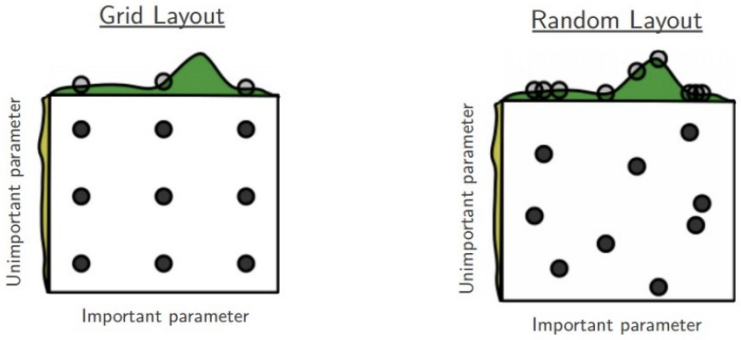


$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}}$$

06 | Discussion

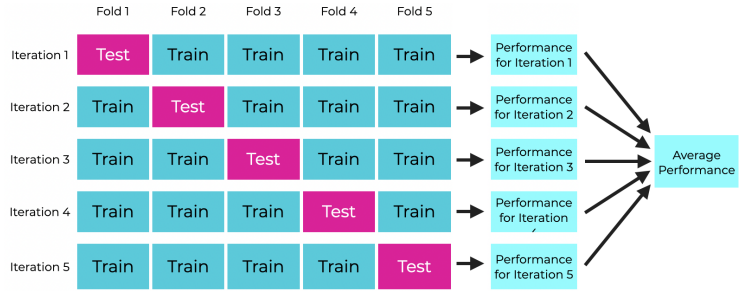
➢ The Necessity of Grid Search

The kNeighborsClassifier module from Scikit-learn, known as KNN, has hyperparameters like n_neighbors, weights, and metric.
Grid search tests combinations of these hyperparameters to find the optimal model performance.



➢ The Necessity of Cross-Validation

Cross-validation helps ensure the model with optimal hyperparameters from grid search doesn't overfit a specific subset and performs well on the entire dataset. It involves splitting the dataset into training and validation sets, evaluating model performance, and repeatedly reassessing the model on different splits of the training set, thus preventing overfitting.



As a result, the optimal hyperparameters obtained through these two processes are {n_neighbors = 1, 'metric': 'manhattan', 'weights': 'uniform'}, and **the accuracy is 0.92**.

Reference

1. OpenAI. (2024). ChatGPT (4) [Large language model]. https://chat.openai.com
2. Lecture Notes on Module 6-5-1 Classification model evaluation
3. Jungyu Choi, Sungbin Im.(2023).Leak Detection and Classification of Water Pipeline based on SVM using Leakage Noise Magnitude Spectrum.Journal of the Institute of Electronics and Information Engineers,60(2),6-14.
4. Yu, T., Chen, X., Yan, W.J., Xu, Z., & Ye, M. Leak detection in water distribution systems by classifying vibration signals. Mechanical Systems and Signal Processing.
5. 상수관로 누수 감지 데이터, ㈜유솔, AI-HUB (2020)
6. Zhang, Wenhao. (2017). Machine Learning Approaches to Predicting Company Bankruptcy. Journal of Financial Risk Management. 06. 364-374. 10.4236/jfrm.2017.64026.
7. 최준규, 임성빈. (2023). 누수 잡음 크기 스펙트럼을 이용한 SVM 기반의 상수관로 누수 감지 및 분류. 전자공학회논문지, 60(2), 6-14, https://www.educative.io/answers/how-to-create-a-confusion-matrix-without-scikit-learn
8. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html
9. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier
10. https://velog.io/@cleansky/K-Nearest-Neighbor-KNN-%EC%95%8C%EA%B3%A0%EB%A6%AC%EC%A6%98

THANK YOU

Group 6

22000415 Yang Chan
22000689 Jeong Yisak
22100809 Hwang Eunji
22100173 Kim Jaehee

HANDONG GLOBAL UNIVERSITY