IT STEP UNIVERSITY

ANOMALY DETECTION

АНАСТАСІЯ ДЕЙНЕКО

К.Т.Н., ДОЦЕНТ

ANASTASIYA.DEINEKO@GMAIL.COM

Пошук викидів (Outlier Detection) і «новизни» (Novelty Detection)

«Новий об’єкт» — це об’єкт, який відрізняється своїми властивостями від об’єктів (навчального) набору. Але на відміну від викидів, його в самому наборі поки немає.

МЕТОДИ ЗНАХОДЖЕННЯ АНОМАЛІЙ:

1. Статистичні тести:
   - Z-нормалізація, візуальний аналіз даних...

2. Модельні тести:

Ідея дуже проста – будується модель, яка описує дані. Такі методи дієві для визначення новизни, але гірше працюють при пошуку викидів.

3. Ітераційні методи

Методи, які складаються з ітерацій, на кожну кроці видаляється група «особо підозрілих об’єктів».

- Методи на основі кластеризації
- Методи на основі машинного навчання

4. Методи на основі метрик

Основна ідея - в них постулюється існування деякої метрики в просторі об'єктів, яка і допомагає знайти аномалію. Інтуїтивно зрозуміло, що у викидів мало "сусідів", а у типової точки багато.

- метрика для знаходження аномалій - мірою аномалій може бути відстань до к-го сусіда (**Local Outlier Factor**)

ОСНОВНІ ВИДИ АНОМАЛІЙ

- Точкові аномалії;
- Контекстуальні аномалії;
- Колективні аномалії.

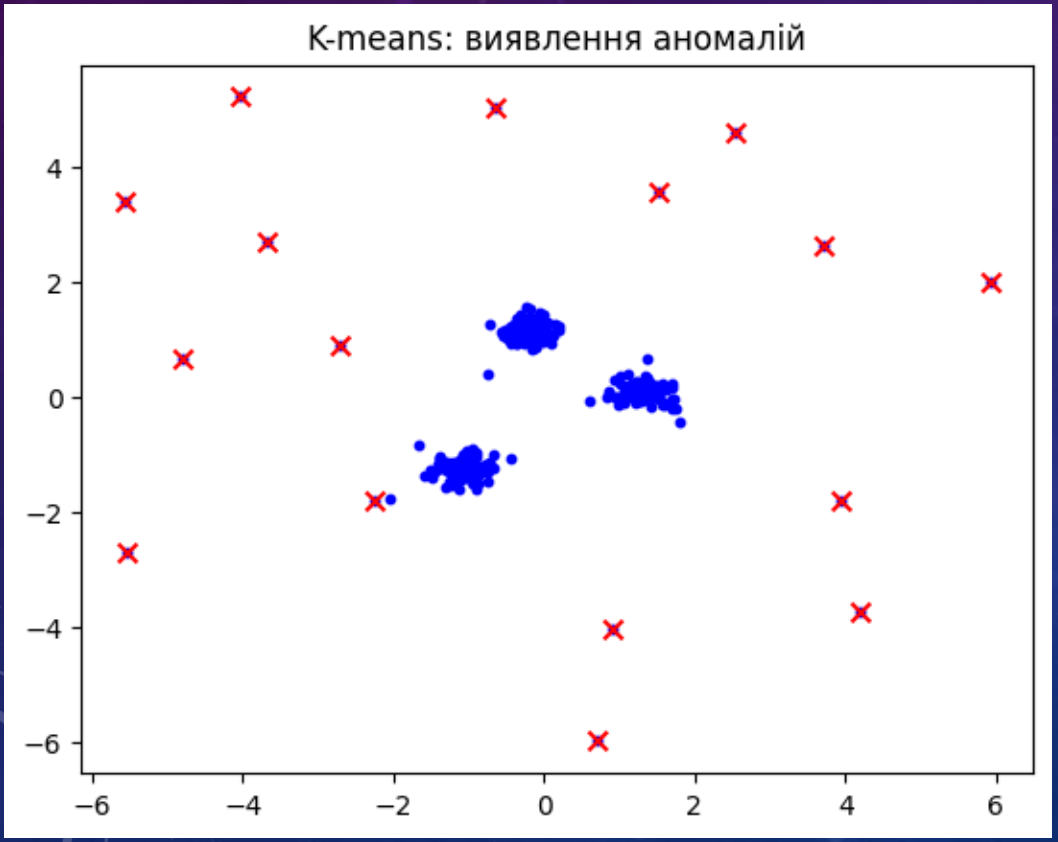
ПРОБЛЕМИ ПРИ ВИЯВЛЕННІ АНОМАЛІЙ

- Невизначеність у природі аномалій;
- Висока розмірність даних;
- Шум у даних;
- Недостатність міток.

МЕТОДИ КЛАСТЕРИЗАЦІЇ ТА ЗМЕНШЕННЯ РОЗМІРНОСТІ

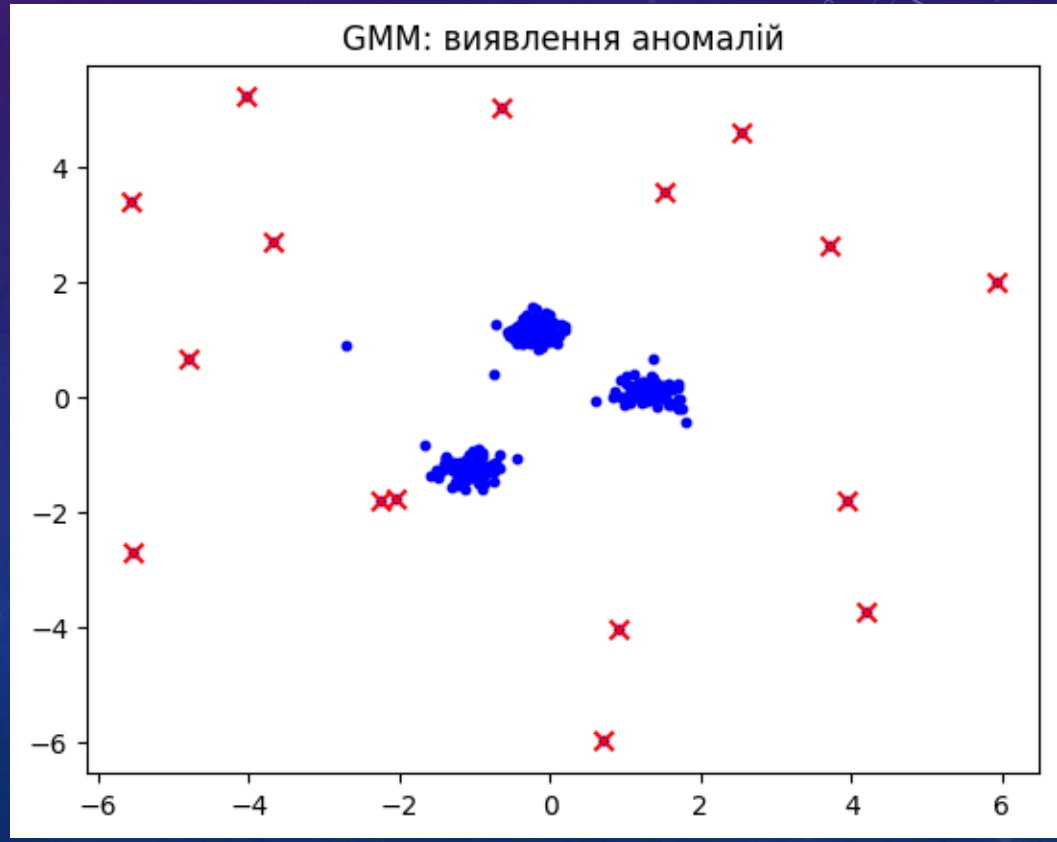
K-means для виявлення аномалій:

- Точки, що знаходяться далеко від центроїдів, можуть розглядатися як аномалії, оскільки вони не належать до жодного типового кластеру.



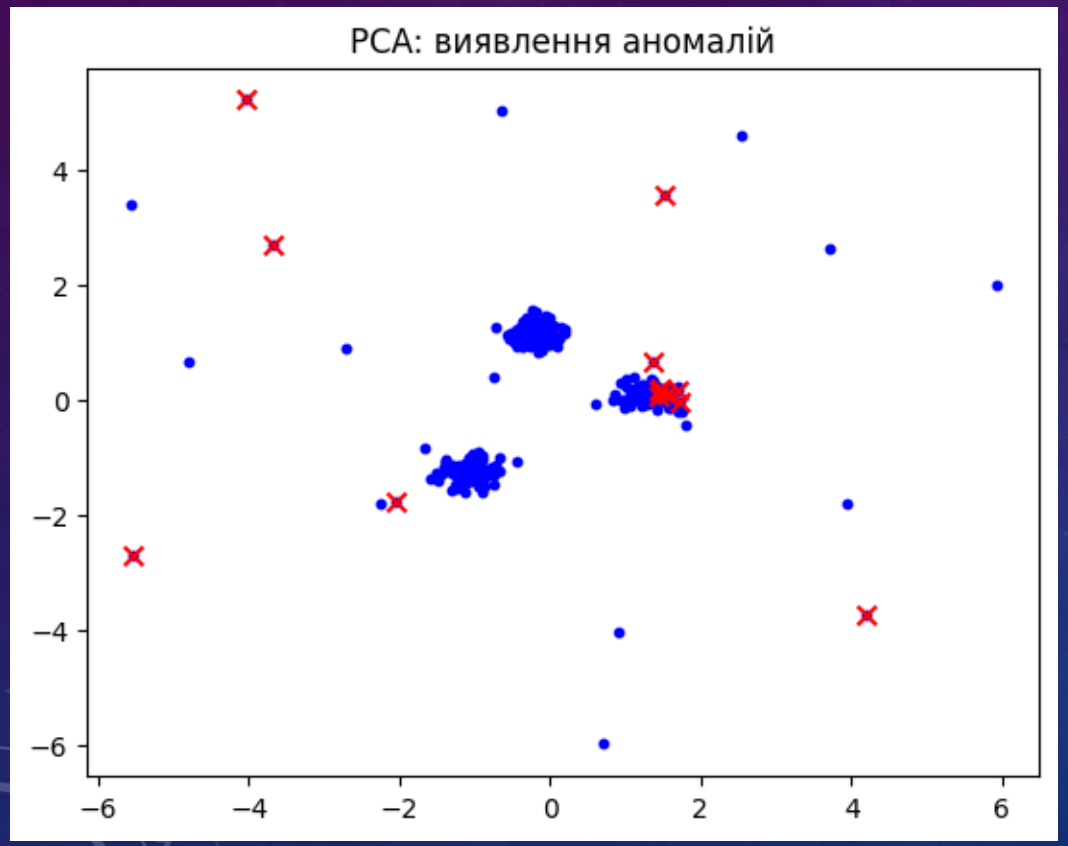
Gaussian Mixture Model (GMM):

- Аномалії — це точки з низькою ймовірністю приналежності до будь-якого кластеру.



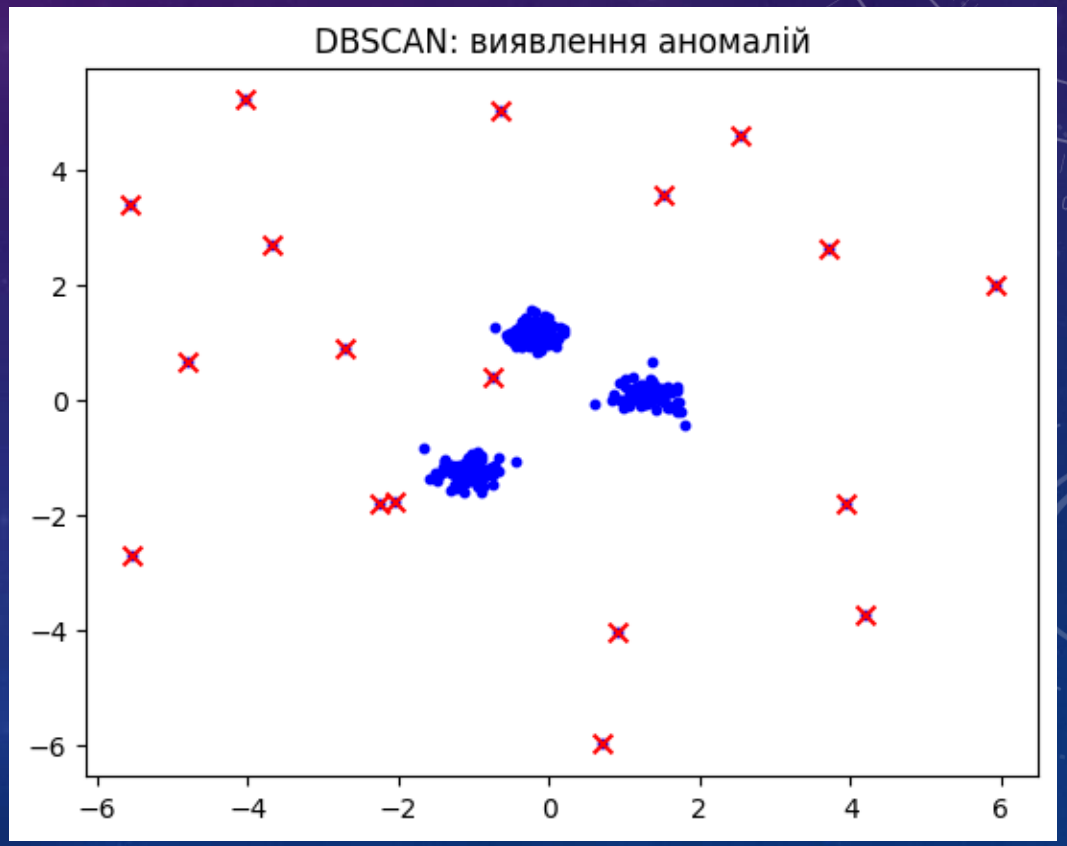
Principal Component Analysis (PCA):

- Точки, які сильно відхиляються від цих нових осей (компонент), можуть розглядатися як аномалії.



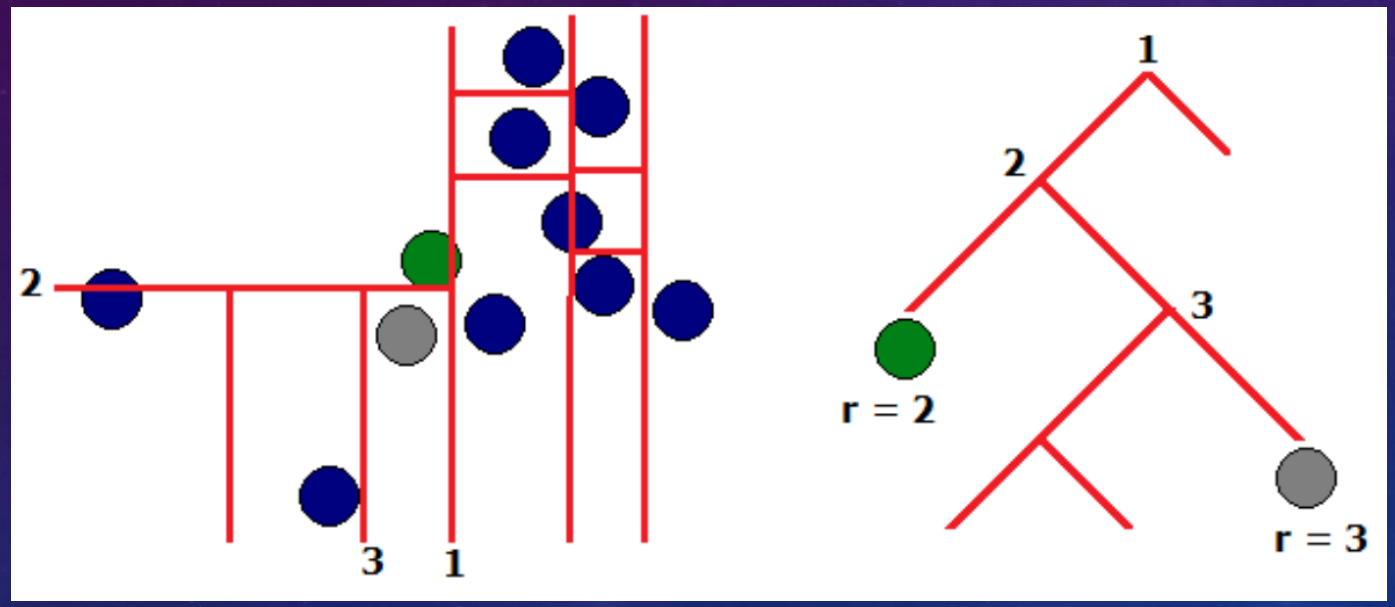
DBSCAN (Density-based Spatial Clustering of Applications with Noise):

- Точки, що залишилися поза кластерами (шумові точки), і є потенційними аномаліями.



Ізолюючий ліс (Isolation Forest):

• Складається з дерев;
• Кожне дерево будується до вичерпання вибірки;
• Для побудови розгалуження в дереві: вибирається випадкова ознака та випадкове розщеплення;
• Для кожного об'єкта міра його нормальності – середнє арифметичне глибин листя, в яке він потрапив (ізолювався).



Важливі параметри реалізації:

n_estimators – кількість дерев;

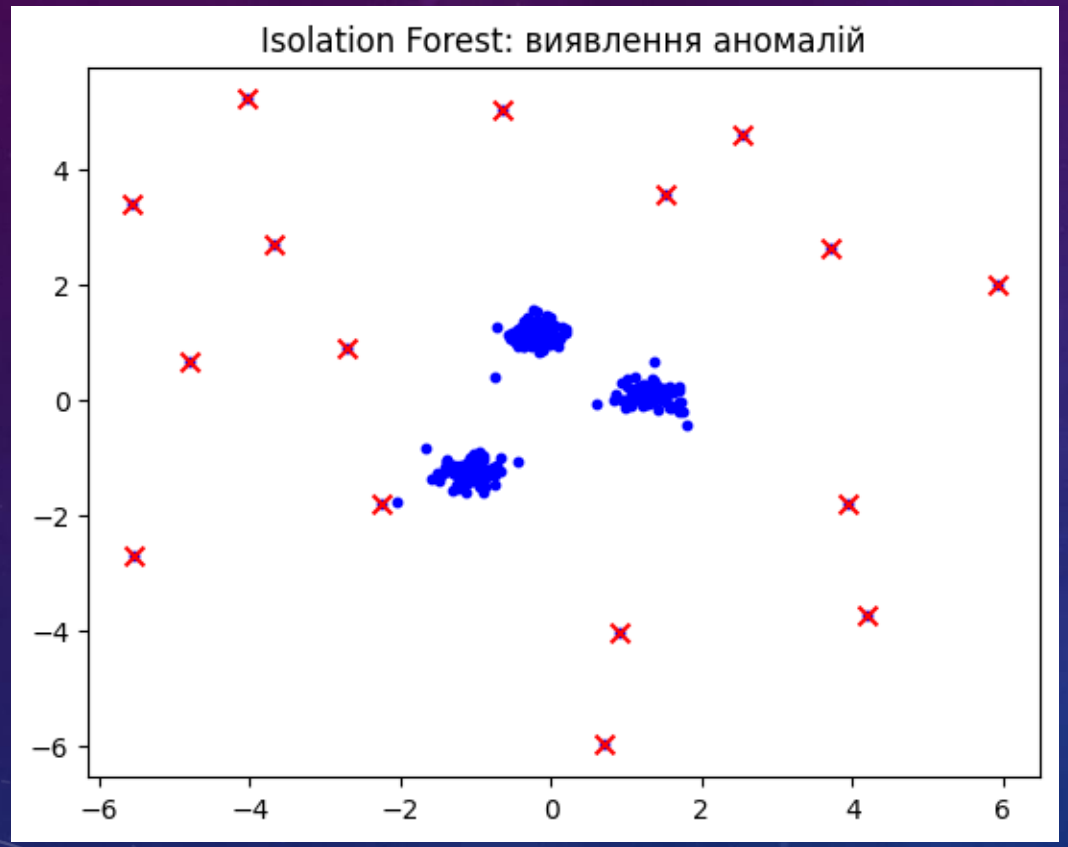
max_samples – обсяг вибірки для побудови одного дерева;

contamination – частка викидів у вибірці (для вибору порога);

max_features – число (або %) ознак, що використовуються при побудові одного дерева;

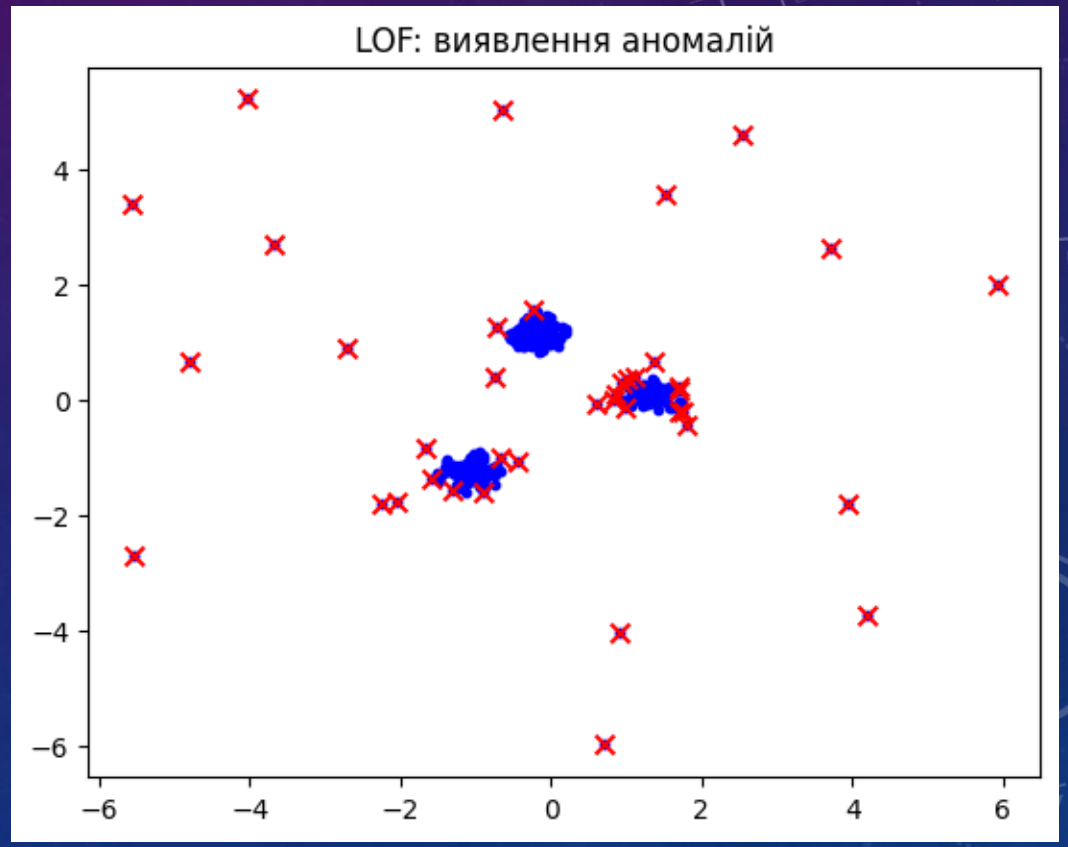
bootstrap – включення режиму бутстрепу для формування підвибірки

Isolation Forest:



Local Outlier Factor (LOF):

- Аномалії — це точки з великим значенням LOF, тобто їх локальна щільність значно нижча за щільність сусідів.



5. Методи машинного навчання:

- Метод опорних векторів для одного класу (OneClassSVM);
- Ізолюючий ліс (IsolationForest);
- Еліпсоідальна апроксимація даних (EllipticEnvelope).

Важливі параметри реалізації OneClassSVM **:**

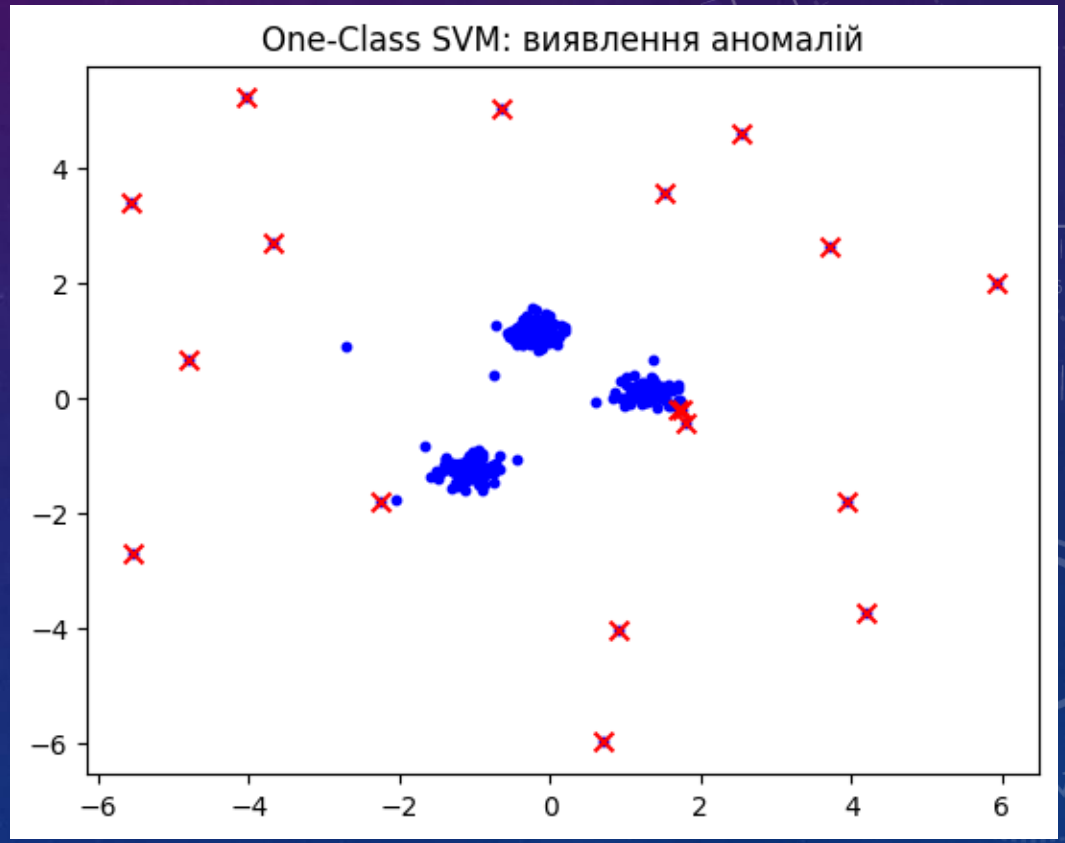
kernel - ядро (лінійне: linear, поліноміальне: poly, радіальні базисні функції: rbf, сигмоїдальне: sigmoid, своє задане);

nu – верхня межа на % помилок та нижня на % опорних векторів;

degree – ступінь для поліноміального ядра;

gamma – коефіцієнт для функції ядра;

coef0 – параметр функції поліноміального або сигмоїдального ядра.



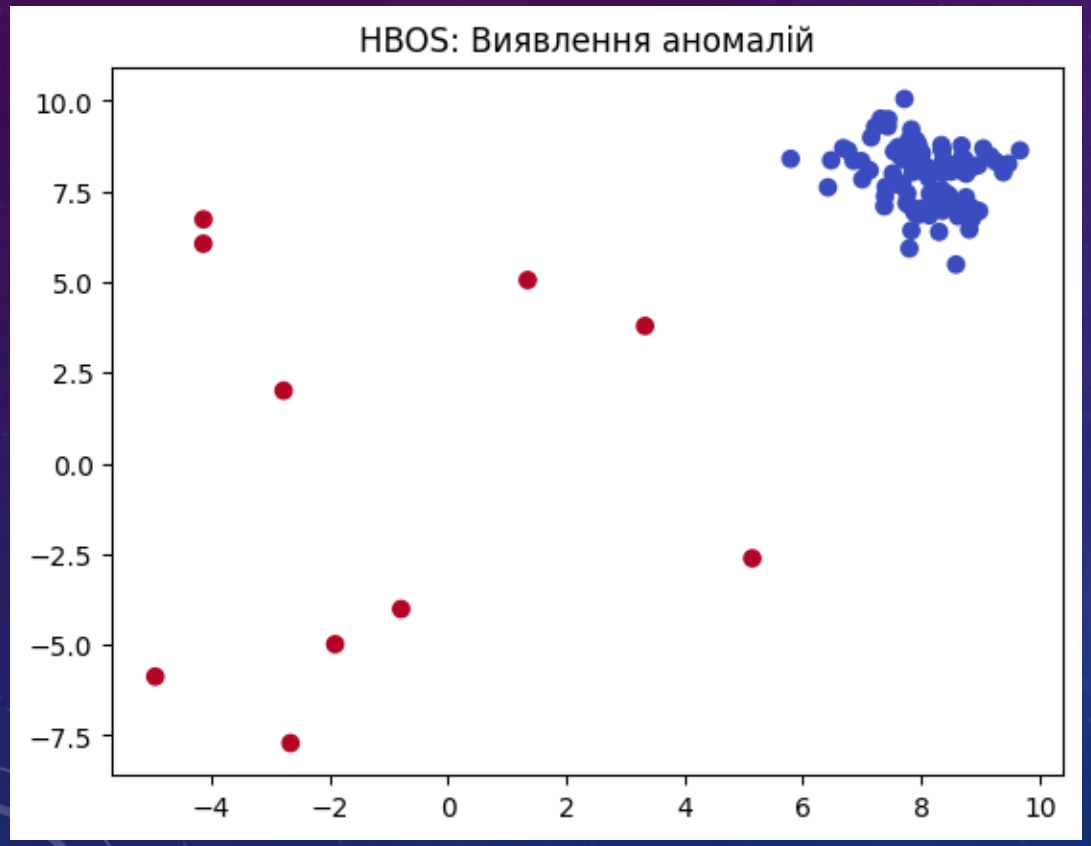
PYOD (PYTHON OUTLIER DETECTION)

- Local Correlation Integral (LCI);
- Histogram-based Outlier Detection (HBOS);
- Angle-based Outlier Detection (ABOD);
- Clustering-Based Local Outlier Factor (CBLOF);
- Minimum Covariance Determinant (MCD);
- Stochastic Outlier Selection (SOS)
- Spectral Clustering for Anomaly Detection (SpectralResidual);
- Feature Bagging;
- Average KNN;
- Connectivity-based Outlier Factor (COF);
- Variational Autoencoder (VAE).

```
pip install pyod
```

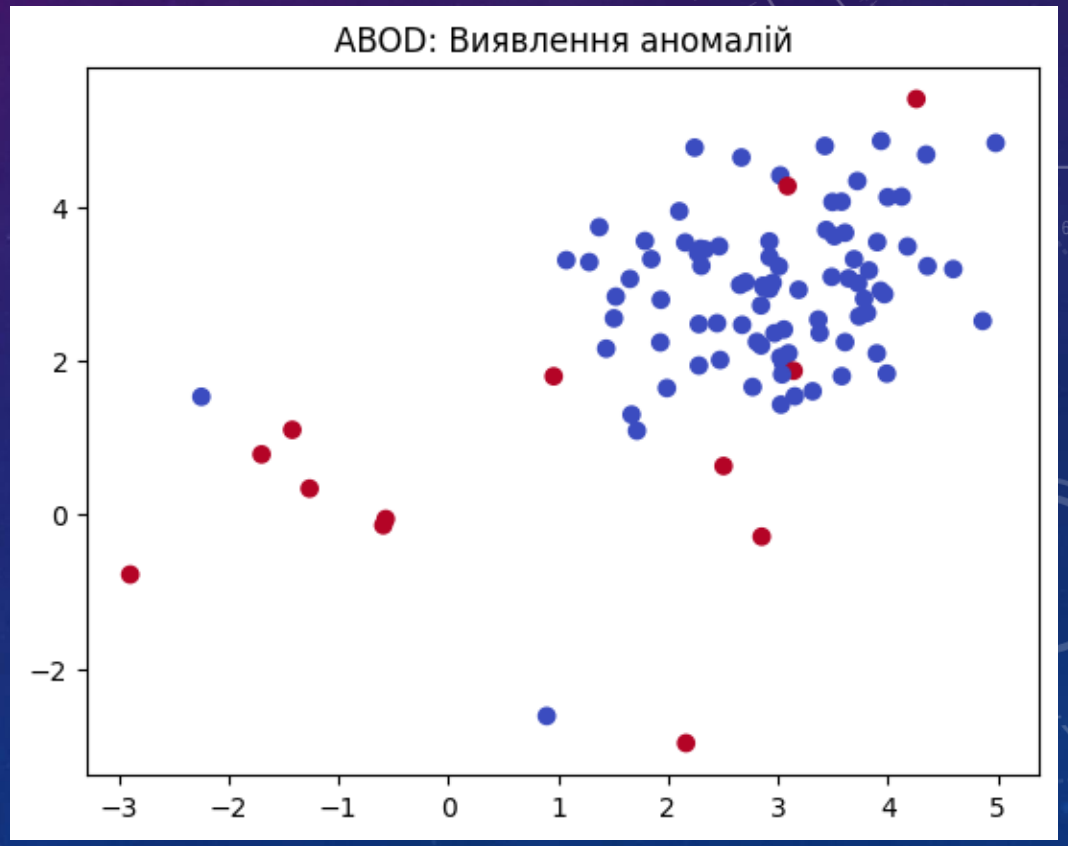
Histogram-based Outlier Detection (HBOS):

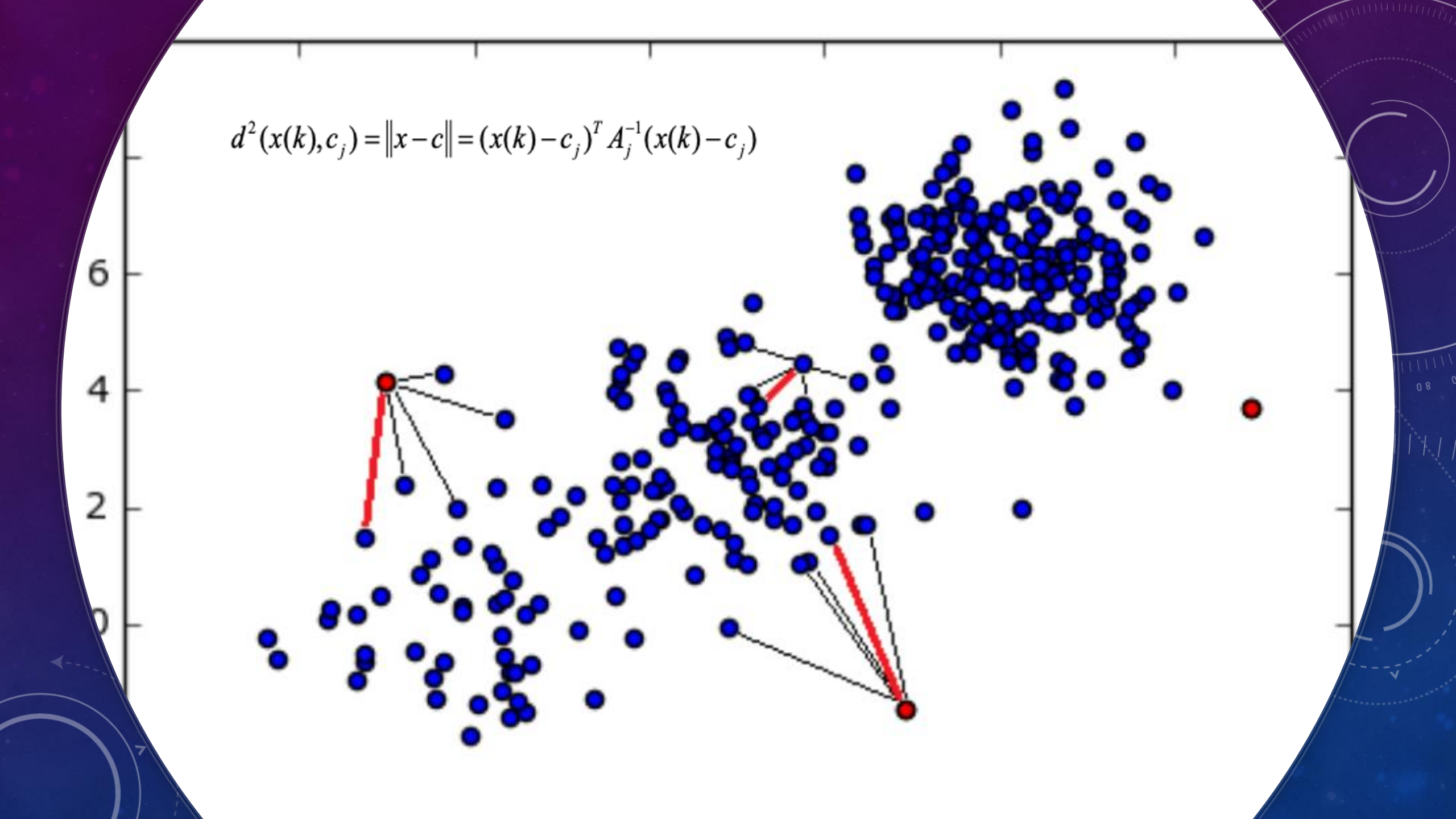
- Аналізує розподіл значень кожної ознаки окремо через гістограми, ідентифікуючи аномалії як ті, що мають низьку ймовірність зустрітись у розподілі.



Angle-based Outlier Detection (ABOD):

- Використовує кути між точками у високовимірному просторі для оцінки того, наскільки ймовірно, що точка є аномальною.




$d^2(x(k),c_j) = \|x-c\| = (x(k)-c_j)^T A_j^{-1}(x(k)-c_j)$
6
4
2
0