Лабораторна робота №1

Методи видобування даних: класифікація, регресія та кластеризація

Опис:

У цій лабораторній роботі потрібно розглянути практичне застосування методів видобування даних, таких як класифікація, регресія та кластеризація. Студенти будуть використовувати процес KDD (Knowledge Discovery in Databases) для збору, обробки та аналізу даних із заданої тематичної області, а також ознайомляться з інструментами для обробки даних (Python, R або SQL).

Мета роботи:

1. Ознайомити студентів з методами видобування даних (класифікація, регресія, кластеризація).
2. Навчити використовувати процес KDD для аналізу даних.
3. Набути практичних навичок роботи з інструментами обробки даних (Python, SQL).
4. Продемонструвати приклади застосування методів у реальних кейсах (маркетинг, фінанси).

Інструменти:

- **Python:** pandas, scikit-learn, matplotlib, seaborn.
- **SQL:** (опціонально для маніпуляцій із базами даних).
- **R:** (як альтернатива для роботи з регресією та кластеризацією).
- **Додаткові джерела:** Kaggle, OpenML для завантаження наборів даних.

Порядок виконання роботи:

1. **Завантаження та підготовка набору даних**
    - **Крок 1:** Завантажте реальний набір даних із відкритих джерел (наприклад, з Kaggle або інших публічних джерел).
        - Приклад: Використайте набір даних про фінансові операції або клієнтські покупки в маркетингу.
    - **Крок 2:** Виконайте попередню обробку даних:
        - Видаліть пропущені або дубльовані дані.
        - Нормалізуйте числові дані.

    Код приклад (Python):

```
import pandas as pd
df = pd.read_csv('dataset.csv')
df.dropna(inplace=True)
df = df.drop_duplicates()
```

2. **Проведення класифікації**
    - **Крок 1:** Виберіть цільовий стовпець для класифікації (наприклад, покупка клієнта або виявлення шахрайства).
    - **Крок 2:** Навчіть модель класифікації (наприклад, Logistic Regression або Decision Tree).
    - **Крок 3:** Оцініть модель за допомогою метрик (точність, precision, recall, F1-score).

    Код приклад (Python):

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report

X = df.drop('target', axis=1)
y = df['target']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)

model = LogisticRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print(classification_report(y_test, y_pred))
```

3. **Проведення регресії**
    - **Крок 1:** Виберіть числову змінну для прогнозування (наприклад, витрати клієнта або прибуток).
    - **Крок 2:** Навчіть модель регресії (Linear Regression або Random Forest Regressor).
    - **Крок 3:** Оцініть модель за допомогою метрик регресії (MAE, MSE, R^2).

    Код приклад (Python):

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error

model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print('MSE:', mean_squared_error(y_test, y_pred))
```

4. **Проведення кластеризації**
    - **Крок 1:** Використайте алгоритм кластеризації (K-means або DBSCAN) для сегментації клієнтів або операцій.
    - **Крок 2:** Визначте оптимальну кількість кластерів за допомогою методу "ліктя".

    Код приклад (Python):

```
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt

kmeans = KMeans(n_clusters=3)
kmeans.fit(X)
y_kmeans = kmeans.predict(X)
```

```
# Візуалізація кластерів
plt.scatter(X[:, 0], X[:, 1], c=y_kmeans)
plt.show()
```

5. **Оцінка результатів та висновки**
    - **Крок 1:** Порівняйте результати класифікації, регресії та кластеризації.
    - **Крок 2:** Проаналізуйте, які методи дали кращі результати для конкретної задачі.
    - **Крок 3:** Проаналізуйте результати та зробіть висновки щодо продуктивності моделей і можливостей для покращення.

Результати роботи:

- Підготовлений набір даних із виконаною очисткою та нормалізацією.
- Алгоритм попередньої обробки даних.
- Навчені та протестовані моделі для класифікації, регресії та кластеризації.
- Оцінка моделей за допомогою відповідних метрик (точність, MSE, метрики кластеризації).

Звіт повинен включати:

- **Опис задачі та набору даних:** Короткий опис того, які дані було використано.
- **Методи обробки даних:** Алгоритм попередньої обробки даних.
- **Результати класифікації, регресії та кластеризації:** Кожна модель повинна мати свої результати з метриками.