**КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ**

**ІМЕНІ ТАРАСА ШЕВЧЕНКА**

**ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ**

**Кафедра інтелектуальних технологій**

**Самостійна робота №1**

з дисципліни «Інтелектуальний аналіз даних»

Тема роботи: «Ансамблеві методи машинного навчання.Random forest»

**Варіант №5**

Виконала:

студентка групи АнД-31

Радоманова С. П.

Перевірив:

Чолишкіна О.Г.

**Київ – 2025**

**Мета роботи:** Вивчити ансамблеві методи машинного навчання, розробити програмну реалізацію для методу Random forest. Мова Python (R).

Завдання і порядок виконання роботи:

1. Вивчити алгоритм методу random forest.

2. Виконати завдання:

2.1 класифікація для одного з датасетів scikit-learn (<https://scikit> learn.org/1.5/datasets/real\_world.html

2.2 регресія для одного з датасетів scikit-learn (<https://scikit> learn.org/1.5/datasets/real\_world.html

3. Оцінити складність і точність проведеної класифікації.

4. Аналіз отриманих результатів, їх змістовна інтерпретація.

Алгоритм методу random forest – це алгоритм машинного навчання для задач класифікації, його ідея полягає в побудові випадкових дерев з вибірки даних (по суті розробки багатьох випадкових алгоритмів), а потім взяття середнього значення з отриманих прогнозованих значень. Для задач класифікації кількість ознак для формування кожної наступної вершини дерева рахується як m=, де n – загальна кількість ознак вектора x, а для задач регресії – m=n/3.

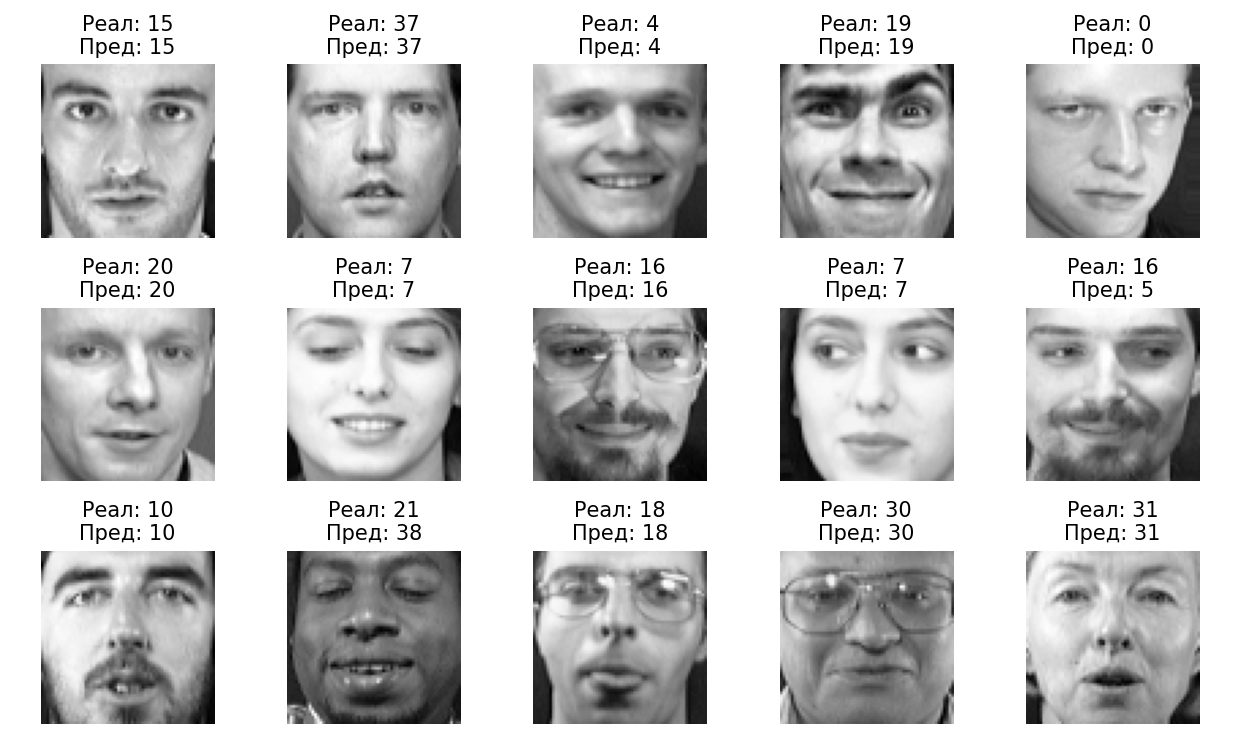
Перевагою такої моделі навчання є те, що зі зростанням кількості дерев алгоритм не перенавчається, в алгоритмі також нема складноналаштовуваних гіперпараметрів.

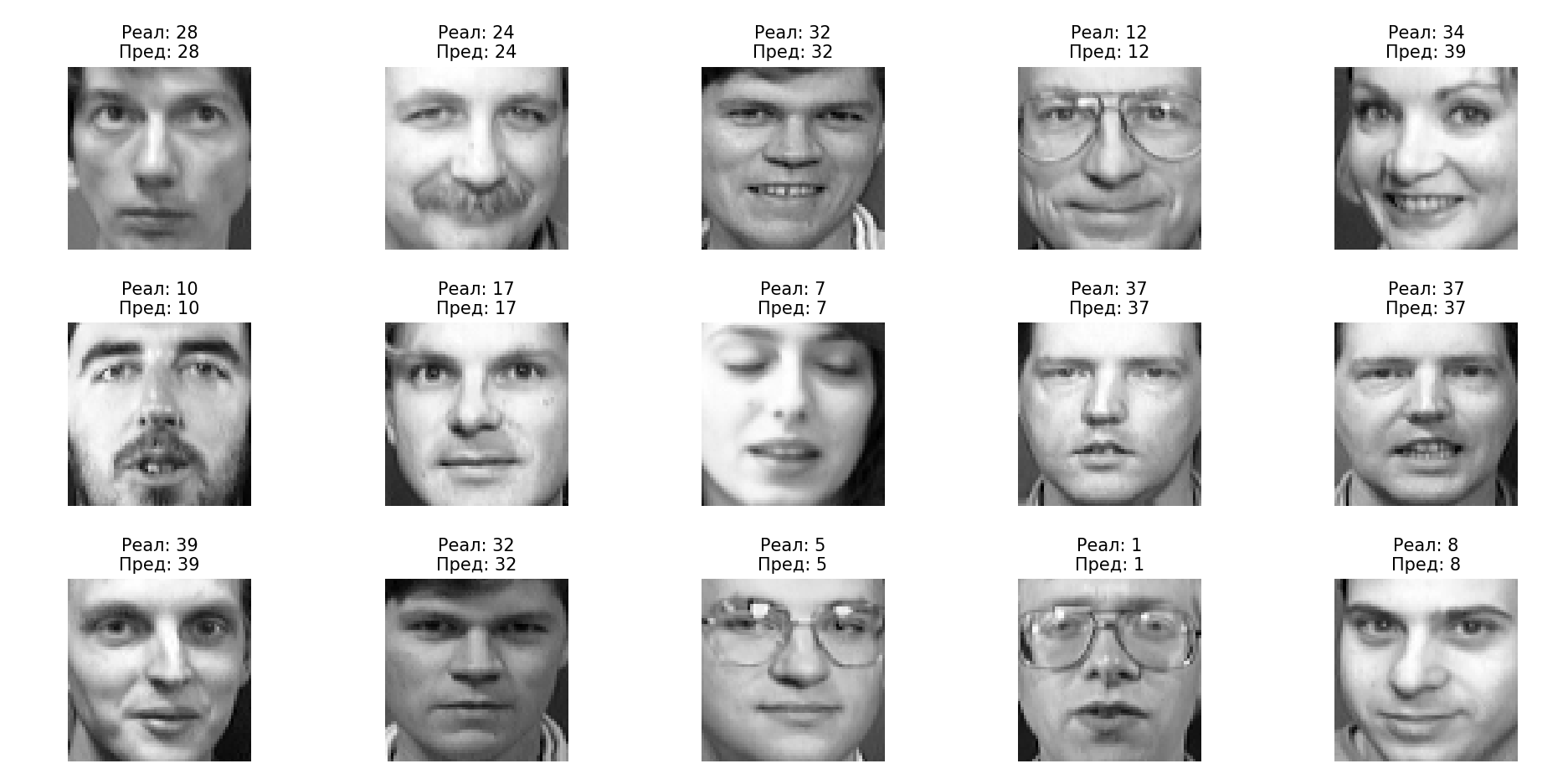
Серед мінусів виділяють складність інтерпретування результатів, відсутність можливості екстраполювання (на відміну від регресії) і використання великих обсягів пам’яті.

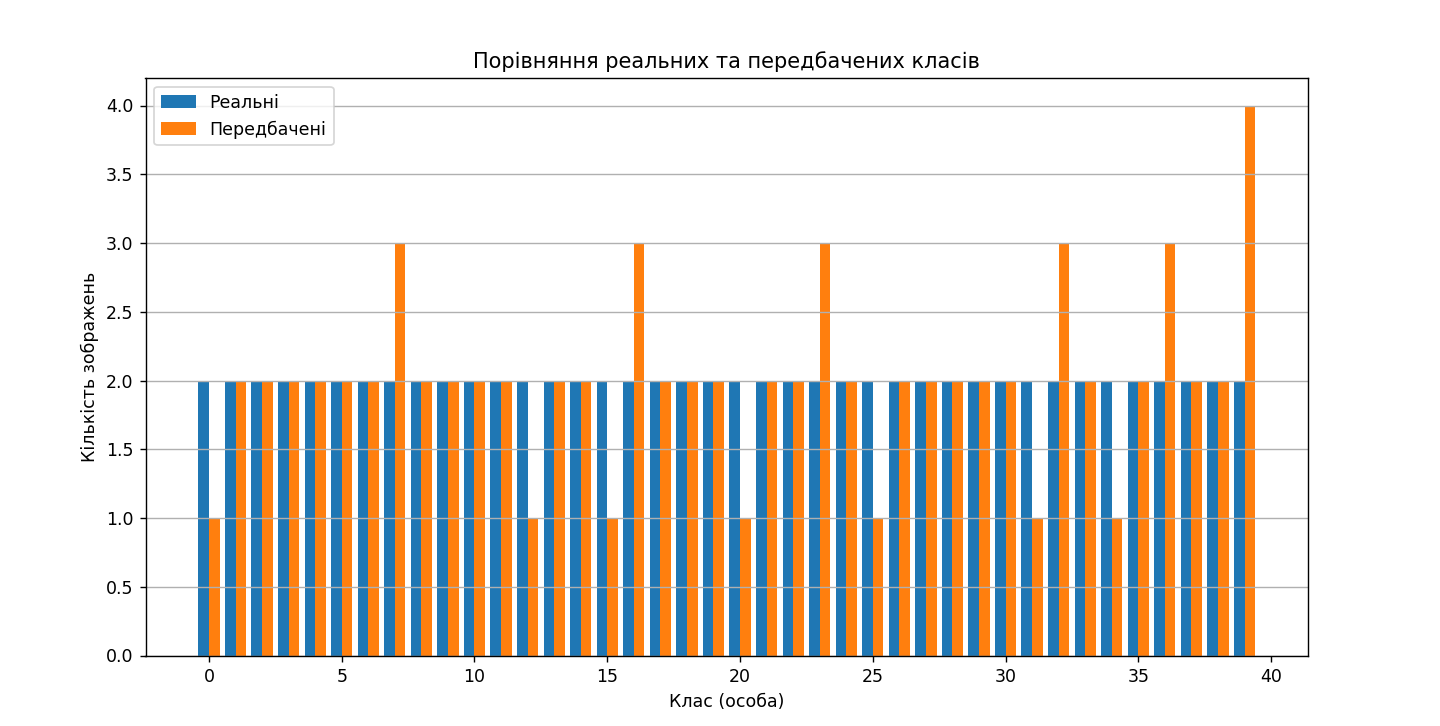
Для класифікації я обрала датасет fetch\_olivetti\_faces. Він імстить 400 зображень (40 людей × 10 фото кожного).Кожне зображення - це 64×64 пікселів у градаціях сірого. Класифікація тут означає навчити модель розпізнавати, чиє це обличчя.

Тут для класифікації використовували функцію з бібліотеки sklearn RandomForestClassifier.

Результати декількох тестів:

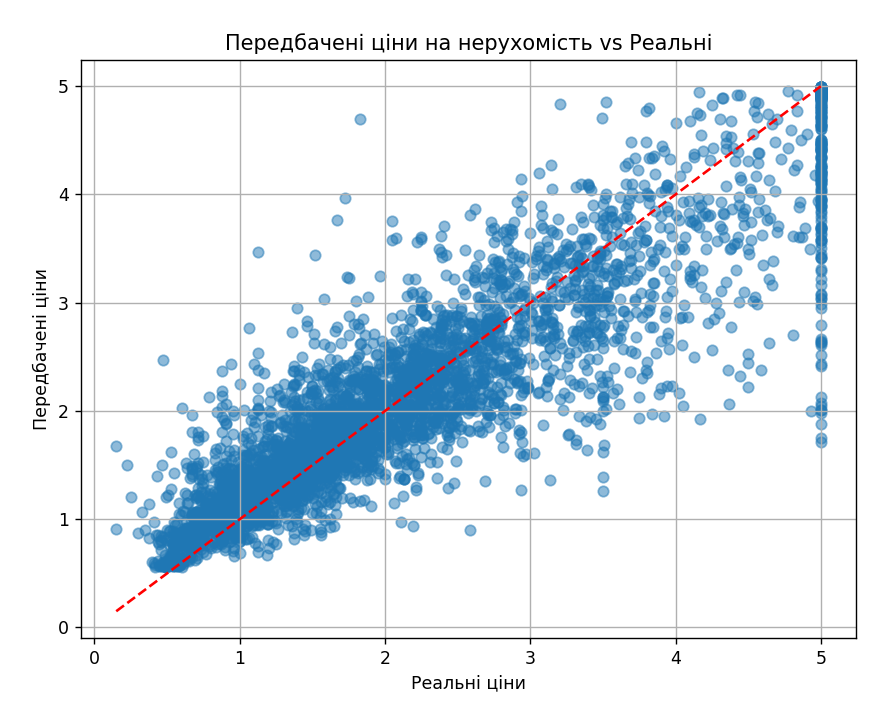






Далі був обраний типовий датасет для задачі регресії - fetch\_california\_housing, що містить дані про нерухомість в різних округах Каліфорнії (8 колонок з ознаками). Задача – передбачити ціну.

Аргументи лишаємо такими самими, 100 дерев глибиною 10.



Було проведено експерименти з ансамблевим методом Random Forest для двох типів задач: класифікації та регресії.

1. Класифікація (fetch\_olivetti\_faces)

Модель будувалася на датасеті з 400 зображень облич (40 класів).

Навчання здійснювалось із використанням 100 дерев глибиною до 10.

Модель змогла розпізнавати обличчя з високою точністю на тестовій вибірці. Це пояснюється тим, що випадкові ліси добре працюють із високорозмірними даними та здатні виокремлювати важливі ознаки навіть при великій кількості пікселів.

Слабкою стороною є те, що модель іноді плутає людей, чиї зображення мають схожі риси або зняті під різними кутами. Це видно у випадках, коли передбачений клас не збігається з реальним.

1. Регресія (fetch\_california\_housing)

Модель застосовувалася для прогнозування вартості житла на основі 8 ознак, що описують район (середній дохід, середній вік будинків, географічні координати тощо).

Random Forest показав здатність відтворювати загальні закономірності та дав адекватні передбачення. На графіку результати здебільшого розташовані близько до ідеальної лінії, що підтверджує добру якість моделі.

Однак, як і у випадку більшості ансамблевих методів, спостерігається певне «згладжування» прогнозів: екстремальні значення цін відтворюються гірше, адже модель орієнтована на середнє по ансамблю.

**Висновок**

У ході роботи було досліджено ансамблевий метод машинного навчання Random Forest, застосований як до задачі класифікації, так і до задачі регресії. Отримані результати показали, що метод є універсальним і добре працює з різними типами даних.

Для задачі класифікації облич модель продемонструвала високу точність і стійкість до перенавчання, що підтверджує ефективність ансамблів дерев для роботи з високорозмірними ознаками. Разом із тим, інтерпретація результатів є складнішою, а схожі об’єкти можуть класифікуватися з помилками.

У задачі регресії метод також дав задовільні результати, дозволивши побудувати адекватну модель прогнозування цін житла. Найкраще він справляється із середніми значеннями, тоді як екстремальні випадки передбачаються менш точно.

Таким чином, можна зробити висновок, що Random Forest є потужним інструментом для інтелектуального аналізу даних, який поєднує простоту налаштувань, високу точність і стійкість. Він особливо корисний у ситуаціях, коли немає необхідності в глибокій інтерпретації результатів, але важлива якість передбачень.