ЗАСТОСУВАННЯ ДЕРЕВ РІШЕНЬ ДЛЯ ВИРІШЕННЯ ЗАДАЧ КЛАСИФІКАЦІЇ НА ПРИКЛАДІ ЗАХВОРЮВАНОСТІ НА ДІАБЕТ

Дудкіна Тетяна Василівна, студентка групи 345а*
Національний аерокосмічний університет ім. М.Є. Жуковського «ХАІ»

Мета класифікації полягає в тому, щоб спрогнозувати мітку класу (Class label), яка являє собою вибір з певного списку можливих варіантів. Класифікація поділяється на бінарну класифікацію (binary classification), яка є окремим випадком поділу на два класи та мультикласову класифікацію (multiclass classification), коли в класифікації бере участь більше двох класів.

Для написання коду для побудови дерева рішень було використано середовище розробки Spyder. Для зчитування данних з таблиці було використано бібліотеку Pandas. Також було використано бібліотеку машинного навчання Scikit-learn, яка містить у собі необхідні функції для роботи з деревами рішень [1-2].

У роботі використано базу даних Pima Indians Diabetes DataBase (рис. 1). Кожен приклад представляє окремих пацієнтів та їхні різні медичні характеристики разом із класифікацією діабету. Кількість екземплярів: 768. Кількість атрибутів – 9 (такі як Pregnancies, PG Concentration, Diastolic BP, Tri Fold Thick, Serum Ins, Body Mass Index, Diabetes Pedigree Function, Age, Diabetes).

	A	B	C	D	E	F	G	H	I
1	Pregnancies	PG Concentration	Diastolic BP	Tri Fold Thick	Serum Ins	BMI	DP Function	Age	Diabetes
2	6	148	72	35	0	33,6	0,627	50	Sick
3	1	85	66	29	0	26,6	0,351	31	Healthy
4	8	183	64	0	0	23,3	0,672	32	Sick
5	1	89	66	23	94	28,1	0,167	21	Healthy
6	0	137	40	35	168	43,1	2,288	33	Sick
7	5	116	74	0	0	25,6	0,201	30	Healthy
8	3	78	50	32	88	31	0,248	26	Sick
9	10	115	0	0	0	35,3	0,134	29	Healthy
10	2	197	70	45	543	30,5	0,158	53	Sick
11	8	125	96	0	0	0	0,232	54	Sick
12	4	110	92	0	0	37,6	0,191	30	Healthy

Рисунок 1 – Pima Indians Diabetes DataBase

Данні було розділено на дві частини – данні для навчання та данні для тестування. Наступним кроком є навчання моделі за допомогою класу DecisionTreeClassifier, з бібліотеці Scikit-learn. Далі робиться прогноз, також необхідно отримати оцінку точності, звіт про класифікацію та матрицю помилок. Останнім кроком є візуалізація дерева рішень, використовуючи функцію export_graphviz з модуля tree. Вона записує файл у форматі .dot, який є форматом текстового файлу, призначеним для опису графіків. Є можливість обрати колір вузлів, щоб виділити клас, який

набрав більшість в кожному вузлі, і передати імена класів та ознак, щоб дерево було правильно розмічено [1-2].

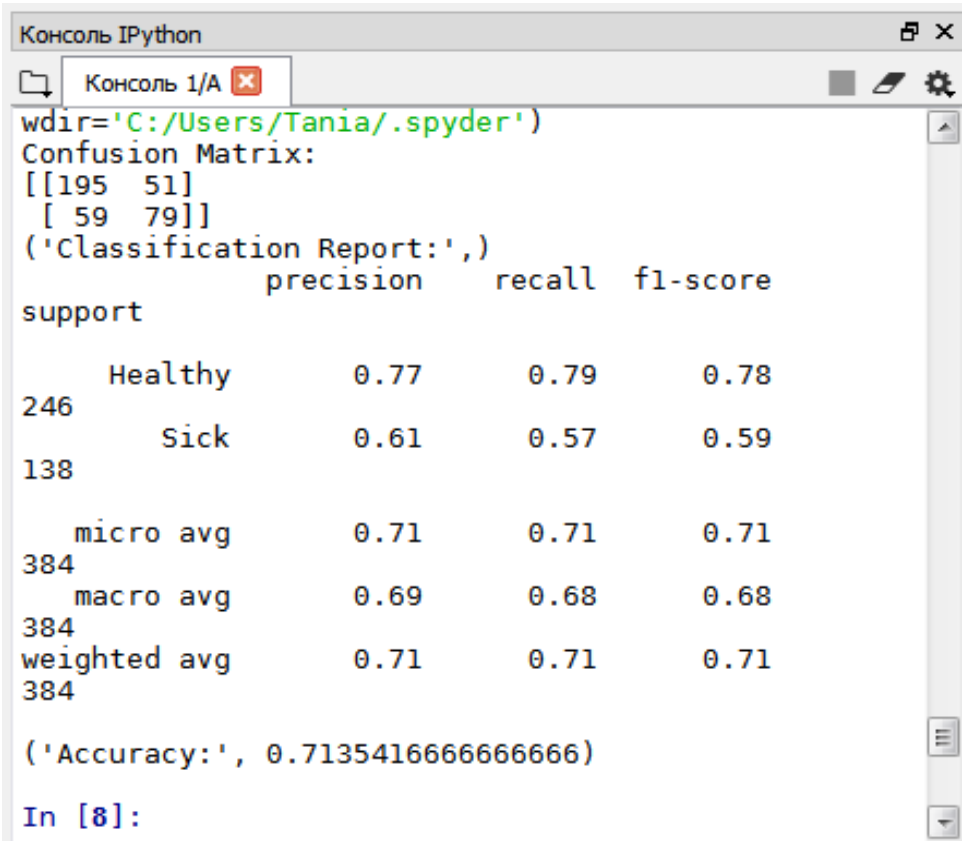

Рисунок 2 – Результати для даних розділених у відношенні 70% та 30%

В ході роботи було проаналізовано три варіанти співвідношень поділу даних: 70% на 30%, 50% на 50%, 30% на 70%. Отримано оцінки точності класифікації: 0.65, 0.71, 0.54 відповідно. Такі результати можливо пояснити проблемою «перенавчання моделі».

Список використаної літератури

1. Рашка С. Python и машинное обучение / пер. с англ. А. В. Логунова. – М.: ДМК Пресс, 2017. – 418 с.: ил.

2. Мюллер, Андреас, Гвидо, Сара. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными. : Пер. с англ. – СПб. : ООО "Альфа-книга", 2017. – 480 с. : ил. – Парал. тит. англ.

**Науковий керівник – Меняйлов Є. С., ст.викл. каф. 304.*