Практическое задание № 16

Для выполнения этого практического задания мы будем использовать данные «diabetes.csv» из практического задания №14.

Задача: построить дерево решений, определяющее наличие сахарного диабета у женщин.

Данные содержат следующие характеристики:

1. Pregnancies – число случаев беременности
2. Glucose – концентрация глюкозы в крови
3. BloodPressure – артериальное диастолическое давление (мм рт. ст.)
4. SkinThickness – толщина кожной складки трехглавой мышцы (мм)
5. Insulin – 2-х часовой сывороточный инсулин
6. BMI – индекс массы тела
7. DiabetesPedigreeFunction – числовой параметр наследственности диабета
8. Age – возраст

Outcome – **целевая переменная**: 1 – наличие заболевания, 0 – отсутствие

1. Загрузите данные в DataFrame с помощью функции `read_csv` библиотеки pandas.

2. Разделите данные на обучающую и тестовую выборки с помощью функции `train_test_split`.

3. Постройте дерево решений с помощью класса `DecisionTreeClassifier с гиперпараметрами по умолчанию`.

4. Отобразите дерево решений с глубиной 2. Опишите процесс принятия решения.

5. Получите информативность признаков. Какие признаки наиболее значимые, какие - наименее?

6. Оцените качество модели с помощью функции `classification_report`.

7. Какая из моделей лучше подходит для диагностики диабета? Линейная SVM-модель (из задания 14) или дерево решений?

8. Подберите оптимальное значение гиперпараметра `max_depth` с помощью поиска по сетке (класс `GridSearchCV`).

9. Обучите модель с оптимальным `max_depth` и оцените результат.

10. *Какая из моделей лучше всего решает поставленную задачу? Как Вы думаете, если уравнять количество наблюдений по классам для обучения модели, качество улучшится или нет?*