

Практическое задание № 1.6.5

Для выполнения этого практического задания мы будем использовать данные «diabetes.csv» из практического задания №1.6.4.

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`.
4. Оцените качество модели с помощью функции `classification_report`.
5. Какая из моделей лучше подходит для диагностики диабета? Линейная SVM-модель (из задания 1.6.4) или дерево решений?
6. Подберите оптимальное значение гиперпараметра `max_depth` с помощью поиска по сетке (класс `GridSearchCV`).
7. Обучите модель с оптимальным `max_depth` и оцените результат.
8. *Какая из моделей лучше всего решает поставленную задачу?*

Подсказка к заданию №6:

```
from sklearn.model_selection import GridSearchCV
parameters = {'max_depth':[2, 5, 10, 20, 100]} # можно указать другие значения
dtc = DecisionTreeClassifier()
grid = GridSearchCV(dtc, parameters)
grid.fit(X_train, y_train)
grid.best_estimator_ # лучшая модель
```