Решение задачи классификации и регрессии с помощью дерева решений

Для работы был выбран датасет [Graduate Admission](https://www.kaggle.com/mohansacharya/graduate-admissions), в котором описывается вероятность поступления в американский университет на основе различных оценок студента, и факте проведения научных исследований.

Датасет имеет следующий вид:

№ студента, ряд оценок, приведенное среднее значение оценок, фокт проведения научных работ, вероятность поступления.

Для постановки задачи классификации был выбран таргет: превышает ли вероятность поступления студента медианного значения.

Для постановки задачи регрессии — вероятность поступления студента

Вектор признаков для задачи классификации и регрессии — общий — значения оценок и факт выполнения исследований. Номер студента не используется, так как не несет в себе информации, которая может нам пригодиться.

Вся выборка разделяется на обучающую и тестовую, в отношении 2/1 соответственно.

Для того, чтобы выбрать лучшие гиперпараметры(критерий информативности, максимальная глубина дерева, минимальная доля для разделения) для алгоритма машинного обучения проводился подбор по сетке, с помошью кроссвалидации на четырех фолдах. В качестве метрики качества возьмем метрику f1 - для классификации, MSE – для регрессии.

Наилучшие параметры для классификатора:

* Критерий информативности — индекс Джини
* Максимальная глубина дерева — 3
* Минимальная доля для разделения — 0.4

Для анализа качества работы классификатора были использованы метрики f1 и accuracy. Полученные значения:

accuracy: 0.824

f1: 0.82197

Самые важные признаки, использованные в классификаторе — CGPA(86%), GRE Score (14%)

Наилучшие параметры для регрессора:

* Критерий информативности — MAE
* Максимальная глубина дерева — 3
* Минимальная доля для разделения — 0.8

Для анализа качества работы регрессора были использованы метрики MSE и MAPE. Полученные значения:

MSE: 0.01

MAPE: 13.4%

Самые важные признаки, использованные в модели — CGPA(100%)

**Заключение.**

В ходе проведенной работы были найдены оптимальные параметры для моделей, решающих задачи классификации и регрессии, а также была получена информация, что самый важный признак, влияющий на вероятность поступления в американский ВУЗ — среднее значение всех оценок