Университет ИТМО

Проект

по дисциплине «Визуализация и моделирование»

Автор: Костылев Иван Михайлович
Поток: 1.1
Группа: K3240
Факультет: ИКТ
Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 г.

Описание датасета

Датасет состоит из данных о студентах, их родителей и оценок, полученных ими по различным предметам.

Всего записей: 1000

Формальное описание

Столбец	Описание	Значения	Формат	Шкала
gender	пол студента	male / female	текст	Качеств номинальная
race/ethnicity	расовая классификация	group A / group B / group C / group D	текст	Качеств номинальная
parental level of education	уровень образования родителей	collegue / school / bachelor's degree / others	текст	Качеств номинальная
lunch	оплата обеда	standart / free/reduced	текст	Качеств номинальная
test preparation	подготовка к тесту	none / completed	текст	Качеств номинальная
math score	оценка по математике	0..100	целое число	Колич относительная
reading score	оценка по чтению	0..100	целое число	Колич относительная
writting score	оценка по письму	0..100	целое число	Колич относительная

Задача машинного обучений

Код с нормализацией, обучением модели и обработкой результатов приведен в Google Colab

https : *//colab.research.google.com/drive/*10*O*50*LdP*0*oLxHryzGaBIZ*97*ue*$_v$*N*9*sharing*

Данные разделим в сотношении 7:3 (обучение : тест)

1. Ранее мы выяснили, что столбцы с данными корелируют друг с другом. Сейчас мы хотим **предсказать оценки студентов на основе личных данных студентов**

Результат:

коэффициент детерминации = 0.15604485344859867

MSE = 14.526800407137248

Такие результаты оказываются достаточно плохими.

2. **Попробуем добавить к нашим данным данные об оценках двух предметов и попробуем предсказать третий**

Результат:

Получается коэффициент детерминации = 0.9435767981339643

MSE = 3.9813127576718337

Наша модель стала предсказывать результаты достаточно точно.

3. **Попробуем предсказать данные о студенте (например, пол) на основе его баллов с помощью k-NN классификатора** *Результат:*

Наиболее точным классификатором оказался классификатор при k=4.

	precision	recall	f1-score	support
0	0.87	0.79	0.83	144
1	0.82	0.89	0.86	156
accuracy			0.84	300
macro avg	0.85	0.84	0.84	300
weighted avg	0.85	0.84	0.84	300

Видно, что мы можем точно предсказать данные о поле студента лишь на основе оценок (без остальных параметров)