НИС

Student Performance

Рыльцева Полина и Колосов Кирилл

10 ноября 2023

Предобработка данных

В процессе предобработки данных мы обработали данные, очистили датасет от нулевых строк и удалили повторения получили что в датасете 639 строк и 39 столбцов.Далее мы приступили к анализу.

Далее с помощью метода corr() была построена зависимость между всеми параметрами в датасете для релевантного подбора параметров для обучения моделей. Для каждой модели были выбраны столбцы с зависимостью хотя бы более, чем 0.2.

Зависимость между параметрами G1, G2, G3 не была проанализирована, так как эти параметры обозначают соответственно оценку за 1 семестр, оценку за 2 семестр и итоговую, их зависимость в любом случае достаточно высокая (по классическим формулам выставления итоговых оценок). Кроме того, высокая корреляция между параметрами Fedu и Medu, которые являются соответственно уровнем образования матери и уровнем образования отца, не была отражена в моделях, поскольку, на наш взгляд, в этом датасете есть более интересные комбинации параметров.

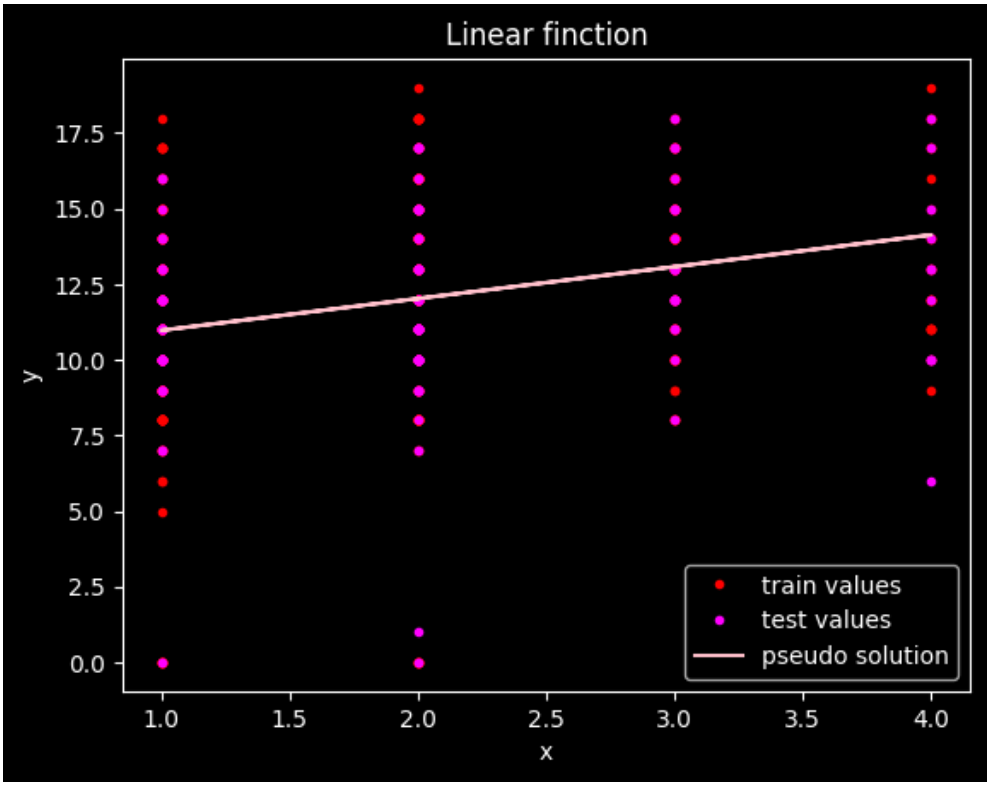
age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
1.000000	-0.107832	-0.121050	0.034490	-0.008415	0.319968	-0.020559	-0.004910	0.112805	0.134768	0.086357	-0.008750	0.149998	-0.174322	-0.107119	-0.106505
-0.107832	1.000000	0.647477	-0.265079	0.097006	-0.172210	0.024421	-0.019686	0.009536	-0.007018	-0.019766	0.004614	-0.008577	0.260472	0.264035	0.240151
-0.121050	0.647477	1.000000	-0.208288	0.050400	-0.165915	0.020256	0.006841	0.027690	0.000061	0.038445	0.044910	0.029859	0.217501	0.225139	0.211800
0.034490	-0.265079	-0.208288	1.000000	-0.063154	0.097730	-0.009521	0.000937	0.057454	0.092824	0.057007	-0.048261	-0.008149	-0.154120	-0.154489	-0.127173
-0.008415	0.097006	0.050400	-0.063154	1.000000	-0.147441	-0.004127	-0.068829	-0.075442	-0.137585	-0.214925	-0.056433	-0.118389	0.260875	0.240498	0.249789
0.319968	-0.172210	-0.165915	0.097730	-0.147441	1.000000	-0.062645	0.108995	0.045078	0.105949	0.082266	0.035588	0.122779	-0.384210	-0.385782	-0.393316
-0.020559	0.024421	0.020256	-0.009521	-0.004127	-0.062645	1.000000	0.129216	0.089707	-0.075767	-0.093511	0.109559	-0.089534	0.048795	0.089588	0.063361
-0.004910	-0.019686	0.006841	0.000937	-0.068829	0.108995	0.129216	1.000000	0.346352	0.109904	0.120244	0.084526	-0.018716	-0.094497	-0.106678	-0.122705
0.112805	0.009536	0.027690	0.057454	-0.075442	0.045078	0.089707	0.346352	1.000000	0.245126	0.388680	-0.015741	0.085374	-0.074053	-0.079469	-0.087641
0.134768	-0.007018	0.000061	0.092824	-0.137585	0.105949	-0.075767	0.109904	0.245126	1.000000	0.616561	0.059067	0.172952	-0.195171	-0.189480	-0.204719
0.086357	-0.019766	0.038445	0.057007	-0.214925	0.082266	-0.093511	0.120244	0.388680	0.616561	1.000000	0.114988	0.156373	-0.155649	-0.164852	-0.176619
-0.008750	0.004614	0.044910	-0.048261	-0.056433	0.035588	0.109559	0.084526	-0.015741	0.059067	0.114988	1.000000	-0.030235	-0.051647	-0.082179	-0.098851
0.149998	-0.008577	0.029859	-0.008149	-0.118389	0.122779	-0.089534	-0.018716	0.085374	0.172952	0.156373	-0.030235	1.000000	-0.147149	-0.124745	-0.091379
-0.174322	0.260472	0.217501	-0.154120	0.260875	-0.384210	0.048795	-0.094497	-0.074053	-0.195171	-0.155649	-0.051647	-0.147149	1.000000	0.864982	0.826387
-0.107119	0.264035	0.225139	-0.154489	0.240498	-0.385782	0.089588	-0.106678	-0.079469	-0.189480	-0.164852	-0.082179	-0.124745	0.864982	1.000000	0.918548
-0.106505	0.240151	0.211800	-0.127173	0.249789	-0.393316	0.063361	-0.122705	-0.087641	-0.204719	-0.176619	-0.098851	-0.091379	0.826387	0.918548	1.000000

Модель 1 (Линейная регрессия)

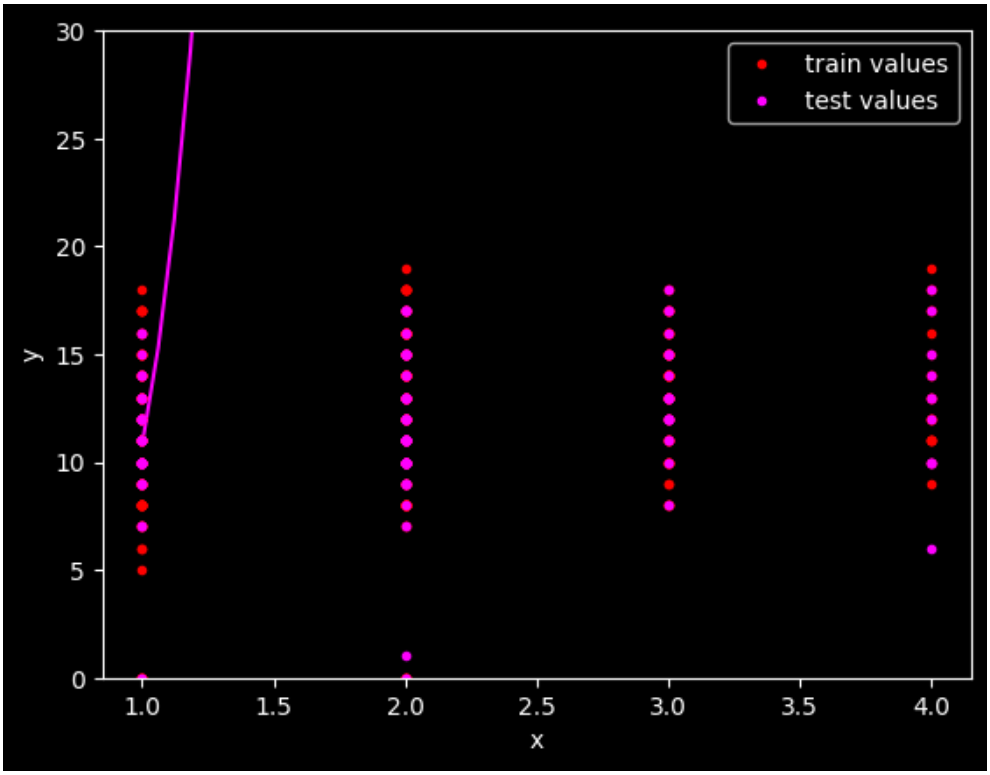
В первой модели будем прогнозировать итоговую оценку за курс (G3) на основании количества времени, проведенного за учебой в неделю - weekly study time, который принимает следующие значения:

- 1 - $<$ 2 hours
- 2 - 2 to 5 hours,
- 3 - 5 to 10 hours,
- 4 - $>$ 10 hours

Коэффициенты для линейной регрессии находятся с помощью метода наименьших квадратов. Была получена следующая зависимость: $1.053309493440232x + 9.919608691690964$
Вывод: Итоговая оценка за курс положительно зависит от количества времени, рассмотрим на графике:



Далее мы попробовали построить нелинейную модель, используя полином 6 степени, однако это не прибавило точности, поскольку функция слишком сильно приближается к распределению оценок при значении параметра studytime = 1. Модель была построена с использованием функции np.polyfit(), которая подбирает полином заданной степени, наилучшим образом приближающий данные:



Модель 2 (Множественная линейная регрессия)

В данной модели мы попробуем предсказать значение параметра go out - going out with friends (numeric: from 1 - very low to 5 - very high) на основе следующих параметров: (numeric:

- Dalc - workday alcohol consumption
- Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
- freetime - free time after school (numeric: from 1 - very low to 5 - very high)
- absences - number of school absences (numeric: from 0 to 93)

Модель строится с использованием библиотечной функции LinearRegression(), в которую передается набор из 4 параметров и значение go out.

$$y = 1.418199098540174 - 0.03468399063350707 \cdot x_1 + 0.34012822960443007 \cdot x_2 +$$
$$+0.3350793416284146 \cdot x_3 + 0.0043975933220913865 \cdot x_4$$

Здесь параметры x_1, x_2, x_3, x_4 перечислены в порядке, соответствующем описанию выше.

Вывод: Мы получили следующий (вроде как вполне логичный) результат: параметр go out положительно зависит от freetime и weekend alchocol consumption в наибольшей степени: weekday alcohol consumption обычно происходит вне дома и в компании с друзьями, а чем больше свободного времени, тем больше возможностей погулять. Кроме того, частота прогулок почти не зависит от числа пропусков школы и daily alcohol consumption

Модель 3 (Задача классификации)

В данной модели попробуем решить задачу классификации путем применения дерева решений. Так как разброс итоговых оценок в целых 20 баллов является слишком большим, преобразуем данные по другим категориям:

- 1 - 10: 1
- 11 - 20: 2

Тогда получится, что итоговые оценки ранжируются по шкале от 1 до 2. Добавим в data столбец result с таким ранжированием, а затем удалим столбы с оценками, которые не потребуются для анализа (в данном случае оставлены параметры - study time и failures

Weekly study time numeric:

- 1 - $<$ 2 hours
- 2 - 2 to 5 hours,
- 3 - 5 to 10 hours,
- 4 - $>$ 10 hours

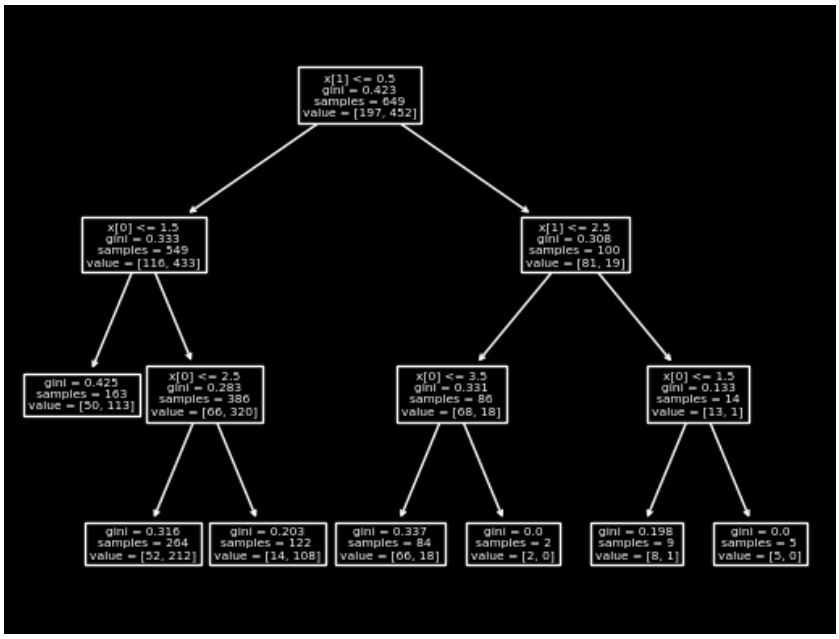
Number of past class failures

- n if $1 \leq$ n$<$3
- else 4

	G1	**G2**	**G3**	**Result**
4	0	11	11	2
2	9	11	11	2
6	12	13	12	2
0	14	14	14	2
0	11	13	13	2
...	...	...	...	...
4	10	11	10	1
4	15	15	16	2
6	11	12	9	1
6	10	10	10	1
4	10	11	11	2

Используя библиотечную функцию tree.DecisionTreeClassifier мы построили дерево, каждая вершина которого является решением в данной модели.

Рассмотрим описанную ранее классификацию:



Рассмотрим на графике: