Университет ИТМО

Практическая работа №5

по дисциплине «Визуализация и моделирование»

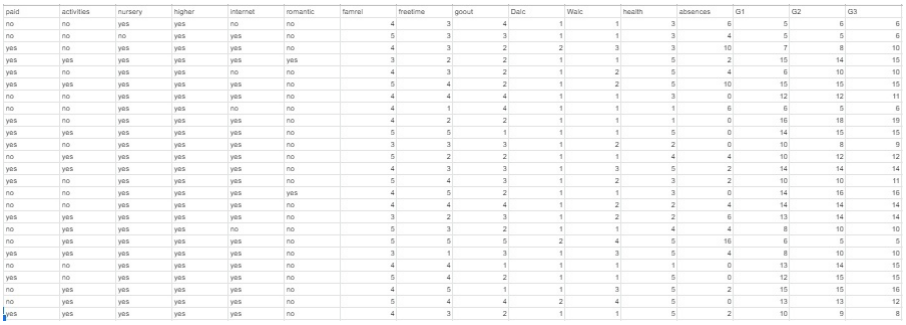
Авторы: Качмазов Артур
Винникова Ева
Новожилова Анна

Поток: 1.2
Группа: К3221
Факультет: ИКТ
Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 г.

1 Описание датасета

В целях работы по дисциплине был выбран датасет, содержащий информацию об учебной успеваемости школьников, употребляющих алкоголь. Датасет состоит из 33 столбцов и содержит порядка 400 записей.



paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
no	no	yes	yes	no	no	4	3	4	1	1	3	6	5	6	6
no	no	no	yes	yes	no	5	3	3	1	1	3	4	5	5	6
yes	no	yes	yes	yes	no	4	3	2	2	3	3	10	7	8	10
yes	yes	yes	yes	yes	yes	3	2	2	1	1	5	2	15	14	15
yes	no	yes	yes	no	no	4	3	2	1	2	5	4	6	10	10
yes	yes	yes	yes	yes	no	5	4	2	1	2	5	10	15	15	15
no	no	yes	yes	yes	no	4	4	4	1	1	3	0	12	12	11
no	no	yes	yes	no	no	4	1	4	1	1	1	6	6	5	6
yes	no	yes	yes	yes	no	4	2	2	1	1	1	0	16	18	19
yes	yes	yes	yes	yes	no	5	5	1	1	1	5	0	14	15	15
yes	no	yes	yes	yes	no	3	3	3	1	2	2	0	10	8	9
no	yes	yes	yes	yes	no	5	2	2	1	1	4	4	10	12	12
yes	yes	yes	yes	yes	no	4	3	3	1	3	5	2	14	14	14
yes	no	yes	yes	yes	no	5	4	3	1	2	3	2	10	10	11
no	no	yes	yes	yes	yes	4	5	2	1	1	3	0	14	16	16
no	no	yes	yes	yes	no	4	4	4	1	2	2	4	14	14	14
yes	yes	yes	yes	yes	no	3	2	3	1	2	2	6	13	14	14
no	yes	yes	yes	no	no	5	3	2	1	1	4	4	8	10	10
no	yes	yes	yes	yes	no	5	5	5	2	4	5	16	6	5	5
yes	yes	yes	yes	yes	no	3	1	3	1	3	5	4	8	10	10
no	no	yes	yes	yes	no	4	4	1	1	1	1	0	13	14	15
yes	no	yes	yes	yes	no	5	4	2	1	1	5	0	12	15	15
no	yes	yes	yes	yes	no	4	5	1	1	3	5	2	15	15	16
no	yes	yes	yes	yes	no	5	4	4	2	4	5	0	13	13	12
yes	yes	yes	yes	yes	no	4	3	2	1	1	5	2	10	9	8

Рис. 1: Датасет часть 1

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no
GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes
GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	yes	no
GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes
GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes
GP	M	16	U	LE3	T	4	3	services	other	reputation	mother	1	2	0	no	yes
GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no	no
GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	yes	yes
GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	no	yes
GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	0	no	yes
GP	F	15	U	GT3	T	4	4	teacher	health	reputation	mother	1	2	0	no	yes
GP	F	15	U	GT3	T	2	1	services	other	reputation	father	3	3	0	no	yes
GP	M	15	U	LE3	T	4	4	health	services	course	father	1	1	0	no	yes
GP	M	15	U	GT3	T	4	3	teacher	other	course	mother	2	2	0	no	yes
GP	M	15	U	GT3	A	2	2	other	other	home	other	1	3	0	no	yes
GP	F	16	U	GT3	T	4	4	health	other	home	mother	1	1	0	no	yes
GP	F	16	U	GT3	T	4	4	services	services	reputation	mother	1	3	0	no	yes
GP	F	16	U	GT3	T	3	3	other	other	reputation	mother	3	2	0	yes	yes
GP	M	17	U	GT3	T	3	2	services	services	course	mother	1	1	3	no	yes
GP	M	16	U	LE3	T	4	3	health	other	home	father	1	1	0	no	no
GP	M	15	U	GT3	T	4	3	teacher	other	reputation	mother	1	2	0	no	no
GP	M	15	U	GT3	T	4	4	health	health	other	father	1	1	0	no	yes
GP	M	16	U	LE3	T	4	2	teacher	other	course	mother	1	2	0	no	no
GP	M	16	U	LE3	T	2	2	other	other	reputation	mother	2	2	0	no	yes
GP	F	15	R	GT3	T	2	4	services	health	course	mother	1	3	0	yes	yes

Рис. 2: Датасет часть 2

Выбор нового датасета обусловлен тем, что используемые нашей командой ранее датасеты (Аниме, К-поп и Мед. Страховка) имели или множество полей, не имеющих явных корреляций(Аниме и К-поп), или корреляции имеются, но столбцов в датасете очень мало для точной работы моделей(Страховка).

В выбранном датасете много полей разных типов, которые легко преобразуются и имеют условно-очевидную корреляцию с конечным результатом.

Гибкость датасета и большой объём различных данных позволяют формулировать и рашать задачи машинного обучения.

2 CRISP-DM

1. Цель

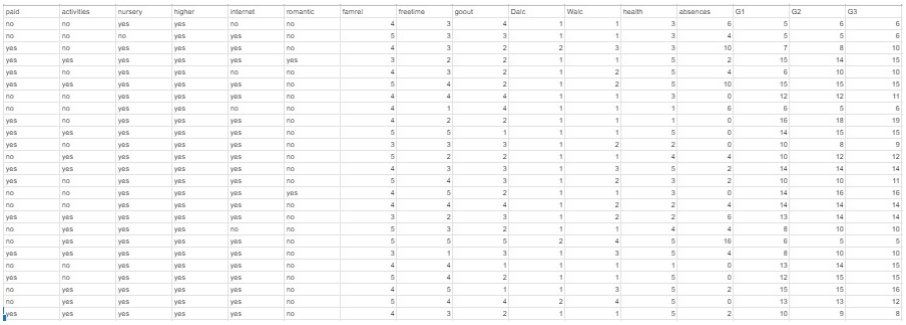
   Цель алгоритма - предсказание итоговой оценки по математике(столбец G3) школьника по данным, приведённым в датасете.

2. Анализ

   При первом взгляде затруднительно выделить определённые данные, которые больше или меньше повлияют на результат. Так как исследование несло социально-психологический характер, данные и предназначены для того, чтобы выделить неочевидные зависимости.

   Единственное, что очевидно - столбцы G1 и G2 не нужны в данной задаче, так как находятся в прямой ассоциации с G3, который мы ищем.

3. Обработка
   Названные столбцы были удалены. Остальные приведены в более удобную форму: строковые поля с определённым набором значений были преобразованы в числовые.



paid	activities	nursery	higher	internet	romantic	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
no	no	yes	yes	no	no	4	3	4	1	1	3	6	5	6	6
no	no	no	yes	yes	no	5	3	3	1	1	3	4	5	5	6
yes	no	yes	yes	yes	no	4	3	2	2	3	3	10	7	8	10
yes	yes	yes	yes	yes	yes	3	2	2	1	1	5	2	15	14	15
yes	no	yes	yes	no	no	4	3	2	1	2	5	4	6	10	10
yes	yes	yes	yes	yes	no	5	4	2	1	2	5	10	15	15	15
no	no	yes	yes	yes	no	4	4	4	1	1	3	0	12	12	11
no	no	yes	yes	no	no	4	1	4	1	1	1	6	6	5	6
yes	no	yes	yes	yes	no	4	2	2	1	1	1	0	16	18	19
yes	yes	yes	yes	yes	no	5	5	1	1	1	5	0	14	15	15
yes	no	yes	yes	yes	no	3	3	3	1	2	2	0	10	8	9
no	yes	yes	yes	yes	no	5	2	2	1	1	4	4	10	12	12
yes	yes	yes	yes	yes	no	4	3	3	1	3	5	2	14	14	14
yes	no	yes	yes	yes	no	5	4	3	1	2	3	2	10	10	11
no	no	yes	yes	yes	yes	4	5	2	1	1	3	0	14	16	16
no	no	yes	yes	yes	no	4	4	4	1	2	2	4	14	14	14
yes	yes	yes	yes	yes	no	3	2	3	1	2	2	6	13	14	14
no	yes	yes	yes	no	no	5	3	2	1	1	4	4	8	10	10
no	yes	yes	yes	yes	no	5	5	5	2	4	5	16	6	5	5
yes	yes	yes	yes	yes	no	3	1	3	1	3	5	4	8	10	10
no	no	yes	yes	yes	no	4	4	1	1	1	1	0	13	14	15
yes	no	yes	yes	yes	no	5	4	2	1	1	5	0	12	15	15
no	yes	yes	yes	yes	no	4	5	1	1	3	5	2	15	15	16
no	yes	yes	yes	yes	no	5	4	4	2	4	5	0	13	13	12
yes	yes	yes	yes	yes	no	4	3	2	1	1	5	2	10	9	8

Рис. 3: До обработки

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	...	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
0	GP	F	18	U	GT3	A	4	4	at_home	teacher	...	4	3	4	1	1	3	6	5	6	6
1	GP	F	17	U	GT3	T	1	1	at_home	other	...	5	3	3	1	1	3	4	5	5	6
2	GP	F	15	U	LE3	T	1	1	at_home	other	...	4	3	2	2	3	3	10	7	8	10
3	GP	F	15	U	GT3	T	4	2	health	services	...	3	2	2	1	1	5	2	15	14	15
4	GP	F	16	U	GT3	T	3	3	other	other	...	4	3	2	1	2	5	4	6	10	10

Рис. 4: После обработки

4. Моделирование Для решения задачи был выбран метод линейной регрессии. Итоговые показатели модели следующие:

   (a) Коэфф. детерминации: 0.27

   (b) Среднеквар. отклонение: 3.14

(c) Точность (Precision): 0.22

(d) Полнота (Recall): 0.16

Был построен график для сравнения тестовых данных и данных построеных на основе машинного обучения. На графике видно, что сохраняется тенденция: участки падения и роста совпадают.

Рис. 5: Сравнение результатов

С учётом небинарности искомой величины, результаты признаны удовлетворительными.

3 Общий вывод

Благодаря методам машинного обучения удалось реализовать алгоритм, позволяющий извлечь практическую выгоду из большого и сложного набора разнообразных данных. Приведённый алгоритм можно усовершенствовать большей выборкой для обучения, исключением помех или выбором другого метода машинного обучения.