«به نام خدا»

تکلیف دوم – سوال دوم – مرضیه علیدادی – 9631983

(کد های مربوط، در دو فرمت py. و ipynb. ضمیمه شده اند. – دیتاست اصلاح شده Diabetes_cleared.csv نیز ضمیمه شده است.)

---

2. با اســـتفاده از قابلیت های مختلف پایتون، مقادیری از ســـتون ها که با تایپ بقیه ی مقادیر آن ســـتون، متفاوت بود را تشخیص دادم. که در واقع این مقادیر نشان دهنده ی NULL بودن آن فیلد ها هستند.

این مقادیر، عبارتند از:

" و NULL و NaN(not a number) و MISS و ?

مثلا در ستون SkinThickness به این ترتیب است:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
252	2	90	80.0	MISS	55	24.4	0.249	24	0
336	0	117	0.0	"	0	33.8	0.932	44	0
422	0	102	64.0	MISS	78	40.6	0.496	21	0
451	2	134	70.0	?	0	28.9	0.542	23	1

حالا با استفاده از مفهوم ستون ها، مشخص می کنم که مقدار 0 در کدام ستون ها نشان دهنده ی NULL است:

ستون BloodPressure که نشان دهنده ی فشار خون است، نمی تواند 0 باشد. پس مقدار 0 در این ستون، نشان دهنده ی NULL بودن آن فیلد است. برای Glucose و SkinThickness و Insulin و BMI هم به همین شکل است.

بقیه ی ستون ها یا دارای مقدار 0 نیستند، و یا مقدار 0 در آن ها معنی دار است. در نتیجه در آن ها به عنوان NULL در نظر گرفته نمی شود.

مثلا بخشی از ستون BMI به این ترتیب است(که این مقادیر 0، نشان دهنده ی missing بودن دیتا در آن فیلد هاست.):

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
9	8	125	96.0	0	0	0.0	0.232	54	1
49	7	105	0.0	0	0	0.0	0.305	24	0
60	2	84	0.0	0	0	0.0	0.304	21	0
81	2	74	0.0	0	0	0.0	0.102	22	0
145	0	102	75.0	23	0	0.0	0.572	21	0
371	0	118	64.0	23	89	0.0	1.731	21	0
426	0	94	0.0	0	0	0.0	0.256	25	0
494	3	80	0.0	0	0	0.0	0.174	22	0

در کتابخانه ی pandas برای مقادیر مفقود، از کلمه ی کلیدی NaN استفاده می شود. و وقتی dataset را به دیتافریتم pandas تبدیل کنیم، مقادیر ذخیره شده با کلمه ی کلیدی NA و NULL، به صورت خودکار، NaN می شوند.

پس من بقیه ی موارد را هم که در بالا ذکر شده بود که نشان دهنده ی مفقود بودن مقدار است، با استفاده قابلیت های موجود، به NaN تبدیل کردم و در یک دیتاست اصلاح شده قرار دادم. و یک نسخه از آن با فرمت csv. نیز تهیه کردم. (از این نسخه ی اصلاح شده، در سوال بعدی استفاده می کنم. - Diabetes_cleared.csv)

بخشی از دیتاست اصلاح شده، که ضمیمه شده است(فیلد های حاوی NaN به صورت خالی در فایل csv. ظاهر شدند.):

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35		33.6	0.627	50	1
1	85	66	29		26.6	0.351	31	0
8	183	64			23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74			25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115				35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96				0.232	54	1
4	110	92			37.6	0.191	30	0
10	168	74			38	0.537	34	1

برای شمردن تعداد NaN های هر ستون، دستوری وجود دارد. و من با استفاده از آن، تعداد NaN های هر ستون را به این ترتیب محاسبه کردم:

```
Pregnancies                 0
Glucose                    20
BloodPressure              26
SkinThickness             146
Insulin                   243
BMI                        10
DiabetesPedigreeFunction    1
Age                         5
Outcome                     5
dtype: int64
```

0 ← Pregnancies

20 ← Glucose

26 ← BloodPressure

146 ← SkinThickness

243 ← Insulin

10 ← BMI

1 ← DiabetesPedigreeFunction

5 ← Age

5 ← Outcome