«به نام خدا»

تکلیف دوم – سوال سوم – 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.)

.3

a) مقادیر مفقود را با عدد 9999 جایگزین کردم:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.0	9999.0	33.6	0.627	50.0	1.0
1	1	85.0	66.0	29.0	9999.0	26.6	0.351	31.0	0.0
2	8	183.0	64.0	9999.0	9999.0	23.3	0.672	32.0	1.0
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21.0	0.0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	1.0
5	5	116.0	74.0	9999.0	9999.0	25.6	0.201	30.0	0.0
6	3	78.0	50.0	32.0	88.0	31.0	0.248	26.0	1.0
7	10	115.0	9999.0	9999.0	9999.0	35.3	0.134	29.0	0.0
8	2	197.0	70.0	45.0	543.0	30.5	0.158	53.0	1.0
9	8	125.0	96.0	9999.0	9999.0	9999.0	0.232	54.0	1.0
10	4	110.0	92.0	9999.0	9999.0	37.6	0.191	30.0	0.0
11	10	168.0	74.0	9999.0	9999.0	38.0	0.537	34.0	1.0
12	10	139.0	80.0	9999.0	9999.0	27.1	1.441	57.0	0.0
13	1	189.0	60.0	23.0	846.0	30.1	0.398	59.0	1.0
14	5	166.0	72.0	19.0	175.0	25.8	0.587	51.0	1.0

b) اولین مقدار موجود در هر ستون را در نظر گرفتم. و مقادیر مفقود هر ستون را، با آن مقدار جایگزین کردم:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.0	94.0	33.6	0.627	50.0	1.0
1	1	85.0	66.0	29.0	94.0	26.6	0.351	31.0	0.0
2	8	183.0	64.0	35.0	94.0	23.3	0.672	32.0	1.0
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21.0	0.0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33.0	1.0
5	5	116.0	74.0	35.0	94.0	25.6	0.201	30.0	0.0
6	3	78.0	50.0	32.0	88.0	31.0	0.248	26.0	1.0
7	10	115.0	72.0	35.0	94.0	35.3	0.134	29.0	0.0
8	2	197.0	70.0	45.0	543.0	30.5	0.158	53.0	1.0
9	8	125.0	96.0	35.0	94.0	33.6	0.232	54.0	1.0
10	4	110.0	92.0	35.0	94.0	37.6	0.191	30.0	0.0
11	10	168.0	74.0	35.0	94.0	38.0	0.537	34.0	1.0
12	10	139.0	80.0	35.0	94.0	27.1	1.441	57.0	0.0
13	1	189.0	60.0	23.0	846.0	30.1	0.398	59.0	1.0
14	5	166.0	72.0	19.0	175.0	25.8	0.587	51.0	1.0

c) مقادیر مفقود هر ستون را، با میانگین مقادیر آن ستون جایگزین کردم:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.000000	35.000000	163.513514	33.60000	0.627	50.0	1.0
1	1	85.0	66.000000	29.000000	163.513514	26.60000	0.351	31.0	0.0
2	8	183.0	64.000000	28.904494	163.513514	23.30000	0.672	32.0	1.0
3	1	89.0	66.000000	23.000000	94.000000	28.10000	0.167	21.0	0.0
4	0	137.0	40.000000	35.000000	168.000000	43.10000	2.288	33.0	1.0
5	5	116.0	74.000000	28.904494	163.513514	25.60000	0.201	30.0	0.0
6	3	78.0	50.000000	32.000000	88.000000	31.00000	0.248	26.0	1.0
7	10	115.0	72.264706	28.904494	163.513514	35.30000	0.134	29.0	0.0
8	2	197.0	70.000000	45.000000	543.000000	30.50000	0.158	53.0	1.0
9	8	125.0	96.000000	28.904494	163.513514	32.50752	0.232	54.0	1.0
10	4	110.0	92.000000	28.904494	163.513514	37.60000	0.191	30.0	0.0
11	10	168.0	74.000000	28.904494	163.513514	38.00000	0.537	34.0	1.0
12	10	139.0	80.000000	28.904494	163.513514	27.10000	1.441	57.0	0.0
13	1	189.0	60.000000	23.000000	846.000000	30.10000	0.398	59.0	1.0
14	5	166.0	72.000000	19.000000	175.000000	25.80000	0.587	51.0	1.0

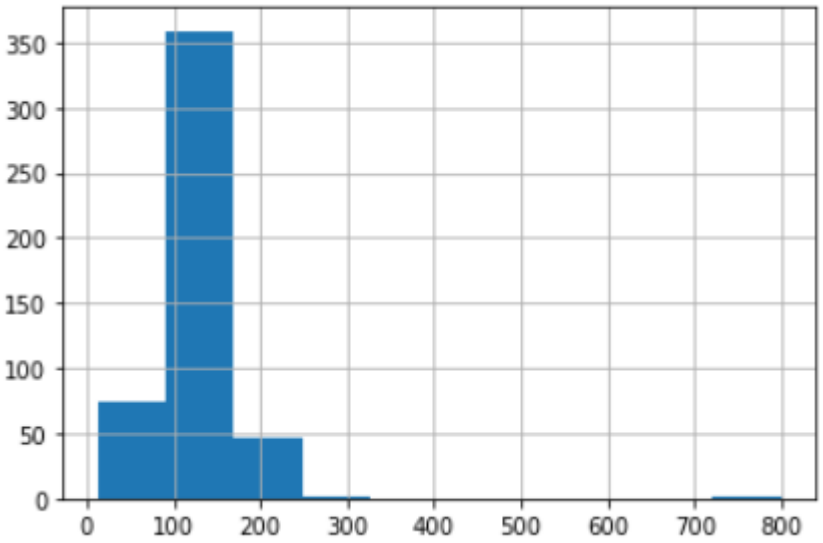
d) مقادیر مفقود هر ستون را، با مد مقادیر آن ستون جایگزین کردم:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.000000	35.000000	163.513514	33.60000	0.627	50.0	1.0
1	1	85.0	66.000000	29.000000	163.513514	26.60000	0.351	31.0	0.0
2	8	183.0	64.000000	28.904494	163.513514	23.30000	0.672	32.0	1.0
3	1	89.0	66.000000	23.000000	94.000000	28.10000	0.167	21.0	0.0
4	0	137.0	40.000000	35.000000	168.000000	43.10000	2.288	33.0	1.0
5	5	116.0	74.000000	28.904494	163.513514	25.60000	0.201	30.0	0.0
6	3	78.0	50.000000	32.000000	88.000000	31.00000	0.248	26.0	1.0
7	10	115.0	72.264706	28.904494	163.513514	35.30000	0.134	29.0	0.0
8	2	197.0	70.000000	45.000000	543.000000	30.50000	0.158	53.0	1.0
9	8	125.0	96.000000	28.904494	163.513514	32.50752	0.232	54.0	1.0
10	4	110.0	92.000000	28.904494	163.513514	37.60000	0.191	30.0	0.0
11	10	168.0	74.000000	28.904494	163.513514	38.00000	0.537	34.0	1.0
12	10	139.0	80.000000	28.904494	163.513514	27.10000	1.441	57.0	0.0
13	1	189.0	60.000000	23.000000	846.000000	30.10000	0.398	59.0	1.0
14	5	166.0	72.000000	19.000000	175.000000	25.80000	0.587	51.0	1.0

e) در روش اول که بدون بررسی، یک عدد ثابت را به مقادیر مفقود نسبت دادم، یک سری داده ی پرت ایجاد شد و پراکندگی داده ها را دچار تغییر ناخواسته کرد. و توجه ما را از جزئیات پراکندگی اصلی، به پراکندگی وسیع تری برد. که این روش اصولاً توصیه نمی شود؛ و بهتر است که عمل جایگزینی را طبق قاعده و تحلیل خاصی انجام دهیم.

در سه روش بعدی که از میانگین، مد و یکی از مقادیر همین ستون، برای جایگزینی استفاده کردیم، توزیع مقادیر در این سطح از تحلیل، تفاوت چندانی با یکدیگر ندارند. و نسبتا نزدیک به توزیع اصلی مقادیر هستند. اما نسبت به روش اول بهتر هستند. به طور کلی، استفاده از یکی از این سه روش، توصیه می شود. که بسته به نوع تحلیل ما، بررسی می شود که کدام یک از این سه روش، بهتر از بقیه عمل خواهد کرد.

توزیع اصلی:



f) در محبث جایگزینی داده های مفقود شـــده، یکی از روش های پرکاربرد رســـیدگی به این مقادیر، در نظر نگرفتن آن هاست. ولی این روش خیلی مناسب نیست.

روش دیگری که پر کاربرد است، استفاده از imputation است. در این روش، داده های مفقود را با یک سری تخمین، با داده های دیگری جایگزین می کنیم. که در این صـــورت، کل داده ها را برای تحلیل در اختیار داریم و از کل آن ها استفاده می کنیم؛ گویی این داده های تخمین زده شده، واقعا همان داده های مشاهده شده هستند.

یک سری روش های معمول، برای این تخمین ها وجود دارد:

1. Mean imputation : میانگین ستون را با استفاده از داده های مشاهده شده ی موجود، محاسبه می کنیم. و این مقدار را جایگزین فیلد های مفقود در آن ســـتون می کنیم. مزیت این روش، این اســـت که، میانگین داده های آن ستون ثابت باقی می ماند. اما این روش مضرات خیلی خیلی زیادی دارد و نسبت به بقیه ی روش ها که در ادامه معرفی می کنم، بدترین است.
2. Substitution : یک نمونه ی جدید از جنس داده هایی که داریم، بررســـی می کنیم. و نتیجه ی مشـــاهدات را جایگزین مقادیر مفقود شده می کنیم.
3. Hot deck imputation : بقیه ی ســمپل های موجود در دیتاســت را در نظر می گیریم. از بین آن ها، آن ســمپل هایی که از نظر بقیه ی متغیر ها، با ســمپل مورد نظر ما که دارای داده ی مفقود اســت، مشــابه اســت را، مد نظر قرار می دهیم. یکی از بین آن ها به صورت تصادفی انتخاب می کنیم و مقدار موجود در آن سمپل که نظیر داده ی مفقود اســت را انتخاب می کنیم و جایگزین داده ی مفقودِ موردنظر می کنیم. یک مزیت این روش این اســت که همواره از داده های معتبر اســـتفاده خواهیم کرد؛ مثلا اگر یک بازه ی مجاز برای متغیری داریم، در این روش، همواره این شرطِ قرار گیری در این بازه، رعایت خواهد شد. مزیت دیگر این است که، استفاده از مولفه ی تصادفی بودن، باعث ایجاد تنوع در داده ها می شود.
4. Cold deck imputation : بقیه ی ســمپل های موجود در دیتاســت را در نظر می گیریم. از بین آن ها، آن ســمپل هایی که از نظر بقیه ی متغیر ها، با ســمپل مورد نظر ما که دارای داده ی مفقود اســت، مشــابه اســت را، مد نظر قرار می دهیم. یکی از بین آن ها با استفاده از روشی سیستماتیک، انتخاب می کنیم و مقدار موجود در آن سمپل

که نظیر داده ی مفقود اســت را انتخاب می کنیم و جایگزین داده ی مفقودِ موردنظر می کنیم. این روش، مشــابه روش قبلی اســت؛ با این تفاوت که به جای انتخاب تصــادفی از بین ســمپل ها، انتخابی ســیســتماتیک خواهیم داشت.

5. Regression imputation : مقدار مفقود شده را برحسب بقیه ی مقادیر حدس می زنیم. در این صورت، میانگین ثابت نمی ماند؛ ولی روابط بین متغیر ها حفظ می شود.
6. Stochastic regression imputation : مثل روش قبلی، مقدار مفقود شده را برحسب بقیه ی مقادیر حدس می زنیم؛ اما اینجا مولفه ی تصــادفی بودن را نیز دخیل می کنیم. این روش، مزایای روش قبلی به علاوه ی مزایای تصادفی بودن را همزمان دارد.
7. Interpolation and extrapolation : یک مقدار برای متغیر مفقود شده، با استفاده از مقادیر بقیه ی متغیر های همان ســمپل، تخمین می‌زنیم. این دو روش، معمولاً فقط در داده های طولی کاربرد دارد. برای اســتفاده از این روش ها، باید احتیاط کرد. و گاهاً قبل از اســتفاده از آن ها، نیاز اســت تا یک ســری پیش فرض های اولیه در نظر گرفته شود.

ما به طور کلی دو رویکرد برای imputation داریم. یا می توانیم از یکی از این 7 روش به تنهایی اســتفاده کنیم؛ و یا اینکه از ترکیبی از آن ها در کنار هم اســتفاده کنیم. مزیت روش اول این اســت که، مفهوم ســاده تری دارد و معمولاً داده‌ی تخمین زده شــده، در بازه ی موردنیاز برای متغیر قرار دارد. و مزیت روش دوم این اســت که، تخمین دقیق تر و بهتری را نتیجه خواهد داد.

برای دیدن منبع، روی این لینک کلیک کنید.

g) با استفاده از متد SimpleImputer مقادیر مفقود شده را با میانگین هر ستون جایگزین کردم. نتیجه مانند زمانی شد که با استفاده از NumPy این کار را کرده بودم:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6.0	148.0	72.000000	35.000000	163.513514	33.60000	0.627	50.0	1.0
1	1.0	85.0	66.000000	29.000000	163.513514	26.60000	0.351	31.0	0.0
2	8.0	183.0	64.000000	28.904494	163.513514	23.30000	0.672	32.0	1.0
3	1.0	89.0	66.000000	23.000000	94.000000	28.10000	0.167	21.0	0.0
4	0.0	137.0	40.000000	35.000000	168.000000	43.10000	2.288	33.0	1.0
5	5.0	116.0	74.000000	28.904494	163.513514	25.60000	0.201	30.0	0.0
6	3.0	78.0	50.000000	32.000000	88.000000	31.00000	0.248	26.0	1.0
7	10.0	115.0	72.264706	28.904494	163.513514	35.30000	0.134	29.0	0.0
8	2.0	197.0	70.000000	45.000000	543.000000	30.50000	0.158	53.0	1.0
9	8.0	125.0	96.000000	28.904494	163.513514	32.50752	0.232	54.0	1.0
10	4.0	110.0	92.000000	28.904494	163.513514	37.60000	0.191	30.0	0.0
11	10.0	168.0	74.000000	28.904494	163.513514	38.00000	0.537	34.0	1.0
12	10.0	139.0	80.000000	28.904494	163.513514	27.10000	1.441	57.0	0.0
13	1.0	189.0	60.000000	23.000000	846.000000	30.10000	0.398	59.0	1.0
14	5.0	166.0	72.000000	19.000000	175.000000	25.80000	0.587	51.0	1.0