**به نام خدا**

**گزارش اول پژوهشی**

**عنوان پژوهش: پیش بینی بیماری دیابت با داشتن فاکتورهای مهم آزمایشگاهی**

* **بررسی و معرفی داده**

داده ها از نوع csv هستند یعنی در هر سطر داده ها با کاما از هم جدا شدند.

برای خواندن داده از نوع csv از pandas و دستور read\_csv استفاده میکنیم.

ابتدا باید pandas را نصب کنیم. پس از آن از دستورهای زیر استفاده می کنیم.

Import pandas as pd

نام pandas را برای راحتی بعد از فراخوانی pd می نامیم.

data = pd.read\_csv('diabetes.csv')

این داده شامل 9 ستون و 768 سطر است.

بخشی از داده:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| preg | plas | pres | skin | test | mass | pedi | age | class |
| 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 8 | 183 | 64 | 0 | 0 | 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 | 0 | 0 | 25.6 | 0.201 | 30 | 0 |
| 3 | 78 | 50 | 32 | 88 | 31 | 0.248 | 26 | 1 |
| 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | 0 |
| 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | 1 |
| 8 | 125 | 96 | 0 | 0 | 0 | 0.232 | 54 | 1 |

چون در فایل csv دیتا عنوان هر ستون را مشخص نکرده است خودمان هنگام خواندن می توانیم عنوان ستونها را بدهیم.

data = pd.read\_csv('diabetes.csv', names=['preg', 'plas', 'pres', 'skin', 'test', 'mass', 'pedi', 'age', 'class'])

این داده ها در واقع اطلاعات مراجعین خانم به بیمارستان در یک سال (سال گذشته) هستند که دیابت نداشتند. اطلاعات هر فرد در ستون یک تا 8 آمده است و سال بعد آن دوباره آزمایش گرفتند و بررسی کردند کدامیک از مراجعین در سال جدید دیابت گرفته اند. اگر دیابت داشتند در ستون آخر ( یعنی class) عدد یک و اگر نداشتند عدد صفر گذاشته اند.

ستون اول (preg): تعداد دفعات بارداری فرد

ستون دوم (plas): پلاسمای خون فرد

ستون سوم (pres): فاکتور فشار خون

ستون چهارم (skin): فاکتور ضخامت پوست

ستون پنجم (test): تست قند خون ( اگر نداده باشد صفر است)

ستون ششم (mass): فاکتور وزن

ستون هفتم (pedi): فاکتور وراثت دیابت ( پدر مادر و ... )

ستون هشتم (age): سن

ستون نهم (class): ستونی است که سال جدید تکمیل شده و بررسی شده این افراد کدامیک دیابت گرفته اند.

در این فایل همه داده ها از نوع عددی هستند. در این داده ها یک سری از ستون ها صفر است که میتواند به این معنی باشد که در آن فرد آن ویژگی بررسی نشده است. (یعنی skin=0 ضخامت پوست صفر است؟؟؟ قطعا نه!)

در پزشکی معمولا وقتی داده ای ندارند به جای خالی گذاشتن آن یک عدد 0 یا 1- می گذارند. این صفرها در ستون test و skin یعنی آزمایش نشده و اگر این موضوع را برطرف نکنیم در پیش بینی به مشکل بر میخوریم. ولی برای ستون preg (یعنی ستون اول) صفر درست است.

پس ما باید اینجا داده ها را درست کنیم.

من با بررسی در اینترنت فهمیدم باید داده ها تمیز یا پاکسازی شوند.

یکی از راههای تمیز کردن داده این است که به جای مقادیر صفر (پوچ) میانگین داده های غیرصفر همان ستون را قرار دهیم.

حالا بررسی می کنیم مثلا ستون skin چه تعداد داده صفر دارد

skines = data[data['skin'] == 0]

با این بررسی فهمیدیم در این فایل 227 ردیف آن دارای skin صفر است

print(len(skines))

DS=data.describe()

print(DS.skin)

با استفاده از دستور data.describe() اطلاعات مفید داده را میبینیم.

اطلاعات زیر مربوط به ستون skin است همانطور که میبینید مقدار min یعنی کمترین مقدار این ستون صفر است.

count 768.000000

mean 20.536458

std 15.952218

min 0.000000

25% 0.000000

50% 23.000000

75% 32.000000

max 99.000000

حالا هر جا ستون skin صفر بود را مقدار میانگین جایگزین می کند.

skin\_mean = data[data['skin'] != 0]['skin'].mean()

data.replace({'skin': 0}, skin\_mean, inplace=True)

حال اگر دوباره describe بگیریم میبینیم مقدار کمترین min در ستون skin تغییر کرده است.

DS\_clean=data.describe()

print(DS\_clean.skin)

count 768.000000

mean 29.153420

std 8.790942

min 7.000000

25% 25.000000

50% 29.153420

75% 32.000000

max 99.000000

یک دستور مهم در بررسی دیتا data.describe() است که برای هر ستون دیتا مقادیر

Count (تعدا ستون) ,

mean (میانگین) ,

std(یعنی انحراف معیار ؟؟؟؟) ,

min(کمترین مقدار هر ستون) ,

25%(؟؟؟؟؟؟) ,

50% (؟؟؟؟؟؟),

75% (؟؟؟؟؟؟),

max (بیشترین مقدار هر ستون)

را می دهد.

برخی از مفاهیم آن را نمی دانم که این جلسه میخواهم در مورد آن بررسی کنم. (این هم خروجی کلی data.describe())

preg plas pres ... pedi age class

count 768.000000 768.000000 768.000000 ... 768.000000 768.000000 768.000000

mean 3.845052 120.894531 69.105469 ... 0.471876 33.240885 0.348958

std 3.369578 31.972618 19.355807 ... 0.331329 11.760232 0.476951

min 0.000000 0.000000 0.000000 ... 0.078000 21.000000 0.000000

25% 1.000000 99.000000 62.000000 ... 0.243750 24.000000 0.000000

50% 3.000000 117.000000 72.000000 ... 0.372500 29.000000 0.000000

75% 6.000000 140.250000 80.000000 ... 0.626250 41.000000 1.000000

max 17.000000 199.000000 122.000000 ... 2.420000 81.000000 1.000000