دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده علوم کامپیوتر

پروژه دوم

تفسیر داده و مصورسازی مجموعه داده‌های افراد مورد حمله قلبی

نگارش
ثمین مهدی پور

استاد راهنما
دکتر قطعی

مهر ۱۴۰۲

چکیده

در این پروژه میخواهیم مجموعه داده ای از افراد با ویژگی‌های زیستی بیولـوژیکی را بـرای کلاس‌بنـدی وقوع حمله قلبی مورد بررسی قرار دهیم. هدف اصلی ما در این پروژه بررسی آماری داده، مصورسـازی و تحلیل آن خواهد بود.

واژه‌های کلیدی:

مصورسازی، بیماری قلبی، داده‌کاوی، تحلیل داده

فهرست مطالب

صفحه

فهرست اشکال

صفحه

No table of figures entries found.

صفحه

فهرست جداول

No table of figures entries found.

فهرست علائم

علائم لاتین

h	ارتفاع
L	طول موج توربولانس
T	پریود توربولانس
U_0	سرعت تعادل وسیله پرنده
$u_{\mathrm{g}}^{\mathrm{B}}$	مولفه سرعت تندباد در راستای محور طولی دستگاه مختصات بدنی نسبت به اینرسی

علائم یونانی

$\Phi(\omega)$	چگالی طیفی قدرت توربولانس
σ	شدت توربولانس
ω	بسامد توربولانس
Ω	بسامد فاصله‌ای

بالانویس‌ها

B	دستگاه مختصات بدنی

زیرنویس‌ها

g	تندباد (گاست)

فصل اول:

مقدمه

بیماری قلبی یکی از شایع ترین بیماری های مزمن در ایالات متحده است که هر سال بر میلیون ها آمریکایی تأثیر می گذارد و بار مالی قابل توجهی بر اقتصاد وارد می کند. تنها در ایالات متحده، بیماری قلبی سالانه جان حدود ۶۴۷۰۰۰ نفر را می گیرد - و آن را به علت اصلی مرگ تبدیل می کند. تجمع پلاک‌ها در داخل شریان‌های کرونری بزرگ‌تر، تغییرات مولکولی مرتبط با افزایش سن، التهاب مزمن، فشار خون بالا و دیابت، همگی از علل و عوامل خطر بیماری قلبی هستند.

در حالی که انواع مختلفی از بیماری عروق وجود دارد، اکثر افراد تنها پس از بروز علائمی مانند درد قفسه سینه، حمله قلبی یا ایست قلبی ناگهانی متوجه می شوند که به این بیماری مبتلا هستند. این واقعیت اهمیت اقدامات پیشگیرانه و آزمایش‌هایی را که می‌توانند به طور دقیق بیماری قلبی را در جمعیت قبل از وقوع پیامدهای منفی مانند انفارکتوس میوکارد (حمله قلبی) پیش‌بینی کنند، برجسته می‌کند.

مراکز کنترل و پیشگیری از بیماری، فشار خون بالا، کلسترول خون بالا و سیگار کشیدن را به عنوان سه عامل خطر اصلی برای بیماری های قلبی شناسایی کرده است. تقریباً نیمی از آمریکایی ها حداقل یکی از این سه عامل خطر را دارند. مؤسسه ملی قلب، ریه و خون به طیف گسترده تری از عوامل مانند سن، محیط و شغل، سابقه خانوادگی و ژنتیک، عادات سبک زندگی، سایر شرایط پزشکی، نژاد یا قومیت و جنس اشاره می کند که پزشکان می توانند در تشخیص بیماری عروق کرونر از آنها استفاده کنند.
تشخیص معمولاً با بررسی اولیه این عوامل خطر رایج به دنبال آزمایش خون و سایر آزمایش‌ها انجام می‌شود. با بررسی این نکات و تحلیل این داده آماری قصد داریم احتمال رخداد بیماری قلبی را در این افراد مورد بررسی قرار دهیم.

.

فصل دوم

تحلیل آماری داده

۲-۱- بررسی ساختار مجموعه داده

برای بررسی ساختار مجموعه داده ابتدا به بررسی سطر و ستون ها میپردازیم.

```
df.shape
```

```
(253680, 22)
```

شکل۲-۱ ابعاد مجموعه داده

ابعاد مجموعه داده شامل ۲۲ ستون و ۲۵۳۶۸۰ سطر است. در گام بعدی به بررسی ستون ها میپردازیم.

مشاهده میشود که مجموعه داده از ستون های زیر که نشان دهنده ویژگی‌های داده هستند تشکیل شده است:

```
df.columns
```

```
Index(['HeartDiseaseorAttack', 'HighBP', 'HighChol', 'CholCheck', 'BMI',
       'Smoker', 'Stroke', 'Diabetes', 'PhysActivity', 'Fruits', 'Veggies',
       'HvyAlcoholConsump', 'AnyHealthcare', 'NoDocbcCost', 'GenHlth',
       'MentHlth', 'PhysHlth', 'DiffWalk', 'Sex', 'Age', 'Education',
       'Income'],
      dtype='object')
```

شکل۲-۲ ستون های مجموعه داده

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 253680 entries, 0 to 253679
Data columns (total 22 columns):
 #   Column                Non-Null Count   Dtype
---  ------                --------------   -----
 0   HeartDiseaseorAttack  253680 non-null  float64
 1   HighBP                253680 non-null  float64
 2   HighChol              253680 non-null  float64
 3   CholCheck             253680 non-null  float64
 4   BMI                   253680 non-null  float64
 5   Smoker                253680 non-null  float64
 6   Stroke                253680 non-null  float64
 7   Diabetes              253680 non-null  float64
 8   PhysActivity          253680 non-null  float64
 9   Fruits                253680 non-null  float64
 10  Veggies               253680 non-null  float64
 11  HvyAlcoholConsump     253680 non-null  float64
 12  AnyHealthcare         253680 non-null  float64
 13  NoDocbcCost           253680 non-null  float64
 14  GenHlth               253680 non-null  float64
 15  MentHlth              253680 non-null  float64
 16  PhysHlth              253680 non-null  float64
 17  DiffWalk              253680 non-null  float64
 18  Sex                   253680 non-null  float64
 19  Age                   253680 non-null  float64
 20  Education             253680 non-null  float64
 21  Income                253680 non-null  float64
dtypes: float64(22)
memory usage: 42.6 MB
```

شکل۲-۳ اطلاعات ستون های مجموعه داده

همانطور که مشاهده میشود ویژگی های مجموعه داده نشان دهنده گروهی از صفات بیولـوژیکی، سـابقه بیماری و شرایط زندگی مورد بررسی در افراد هستند که میخواهیم احتمال رخداد بیماری قلبی را بـرای آنها شناسایی کنیم.

بررسی میکنیم که آیا ستونی مقدار null دارد یا خیر:

```
df.isnull().any()
```

```
HeartDiseaseorAttack    False
HighBP                  False
HighChol                False
CholCheck               False
BMI                     False
Smoker                  False
Stroke                  False
Diabetes                False
PhysActivity            False
Fruits                  False
Veggies                 False
HvyAlcoholConsump       False
AnyHealthcare           False
NoDocbcCost             False
GenHlth                 False
MentHlth                False
PhysHlth                False
DiffWalk                False
Sex                     False
Age                     False
Education               False
Income                  False
dtype: bool
```

شکل۲-۴ بررسی وجود مقدار null در ستون‌های مجموعه داده

همانطور که مشاهده میشود مقدار نامعلومی در ستون ها موجود نیست.

با فراخوانی تابع head ۵ سطر اول داده را مشاهده خواهیم کرد:

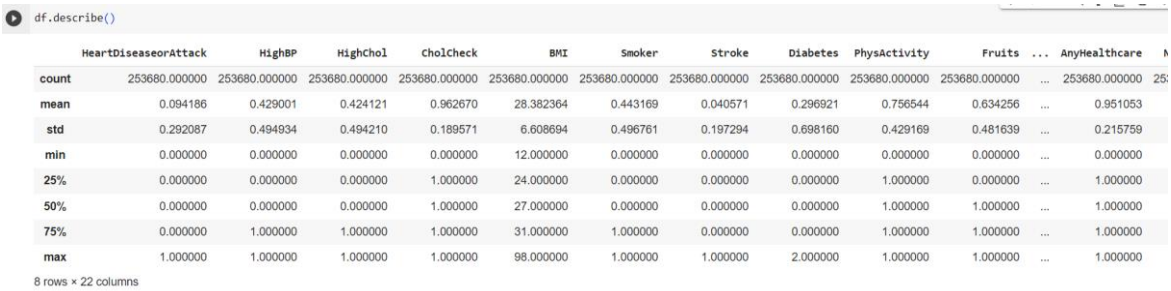
```
[8] df.head()
```

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	A
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	0.0	...	1.0	0.0	5.0	18.0	15.0	1.0	0.0	9
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	0.0	...	0.0	1.0	3.0	0.0	0.0	0.0	0.0	7
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	1.0	...	1.0	1.0	5.0	30.0	30.0	1.0	0.0	9
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	0.0	0.0	0.0	0.0	11
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	1.0	...	1.0	0.0	2.0	3.0	0.0	0.0	0.0	11

5 rows × 22 columns

شکل۲-۵ پنج سطر اول مجموعه داده

در گام بعدی از describe استفاده میکنیم:



```
df.describe()
```

	HeartDiseaseorAttack	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	Diabetes	PhysActivity	Fruits	...	AnyHealthcare	N
count	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	...	253680.000000	25
mean	0.094186	0.429001	0.424121	0.962670	28.382364	0.443169	0.040571	0.296921	0.756544	0.634256	...	0.951053	
std	0.292087	0.494934	0.494210	0.189571	6.608694	0.496761	0.197294	0.698160	0.429169	0.481639	...	0.215759	
min	0.000000	0.000000	0.000000	0.000000	12.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.000000	
25%	0.000000	0.000000	0.000000	1.000000	24.000000	0.000000	0.000000	0.000000	1.000000	0.000000	...	1.000000	
50%	0.000000	0.000000	0.000000	1.000000	27.000000	0.000000	0.000000	0.000000	1.000000	1.000000	...	1.000000	
75%	0.000000	1.000000	1.000000	1.000000	31.000000	1.000000	0.000000	0.000000	1.000000	1.000000	...	1.000000	
max	1.000000	1.000000	1.000000	1.000000	98.000000	1.000000	1.000000	2.000000	1.000000	1.000000	...	1.000000	

8 rows × 22 columns

شکل۲-۶ بررسی آماری ویژگی های مجموعه داده

با استفاده از این فانکشن میتوان به پراکندگی آماری داده ها در هرستون پرداخت. Count نشان دهنده داده های دارای مقدار در هرستون است. Mean میانگین مقدار هرستون را نشان میدهد. Std نشان دهنده انحراف معیار است. min کمترین مقدار هرستون را نشان داده وmax بیشترین مقدار را نشان میدهد. 25%,50%,75% به ترتیب ۲۵،۵۰،۷۵ درصد مقدار داده های هرستون را نشان میدهند که مقدار ۵۰ درصد همان median یا میانه است.

برای تحلیل بهتر داده ها از ازمون های اماری در ادامه استفاده خواهیم کرد:

۲-۲- آزمون های آماری

۲-۲-۱ تی تست

در این آزمون آماری میانگین خروجی داده ها براساس یک ویژگی مقایسه میشود. مثلا در این بخش بر اساس سن افرادی که مورد حمله قلبی قرار گرفته اند را بررسی میکنیم:

```
▾ T-Test

[23] heart_disease_group = df[df['HeartDiseaseorAttack'] == 1]
     healthy_group = df[df['HeartDiseaseorAttack'] == 0]

[24] age_heart_disease_mean = heart_disease_group['Age'].mean()
     age_healthy_mean = healthy_group['Age'].mean()

[25] print("میانگین سن افراد مبتلا به بیماری قلبی: {:.2f} سال".format(age_heart_disease_mean))
     print("میانگین سن افراد سالم: {:.2f} سال".format(age_healthy_mean))

     میانگین سن افراد مبتلا به بیماری قلبی: 40.52 سال
     میانگین سن افراد سالم: 31.26 سال
```

شکل۲-۷ نتیجه تی تست براساس سن

اعمال تی تست بر اساس فعالیت فیزیکی:

based on **Activity**

```
[31] from scipy.stats import ttest_ind

     group_heart_disease = df[df['HeartDiseaseorAttack'] == 1]['PhysActivity']
     group_healthy = df[df['HeartDiseaseorAttack'] == 0]['PhysActivity']

     t_statistic, p_value = ttest_ind(group_heart_disease, group_healthy, equal_var=False)

     print("t-static: {:.2f}".format(t_statistic))
     print("p-value: {:.5f}".format(p_value))

     t-static: -39.75
     p-value: 0.00000
```

شکل۲-۸ نتیجه تی تست براساس فعالیت بدنی

در این بخش داده‌ها به دو گروه تقسیم می‌شوند، یک گروه برای افراد مبتلا به بیماری قلبی (group_heart_disease) و دیگری برای افراد سالم (group_healthy) براساس مقدار ویژگی "HeartDiseaseorAttack".

سپس از تابع ttest_ind از کتابخانه SciPy برای انجام تی-تست بر روی میزان فعالیت بدنی (ویژگی "PhysActivity") بین دو گروه استفاده می‌شود. مقدار equal_var=False برای نشان دادن عدم تساوی واریانس در دو گروه تنظیم می‌شود.

نتایج تی-تست شامل static-t و مقدار p-value می‌باشد. اگر مقدار p-value به اندازه کافی کم باشد (معمولاً کمتر از ۰.۰۵)، نشان‌دهنده وجود تفاوت معنی‌دار در میزان فعالیت بدنی بین دو گروه است.

در نهایت، اگر مقدار p-value کمتر از ۰.۰۵ باشد، می‌توان نتیجه گرفت که میزان فعالیت بدنی ممکن است تأثیری در تعیین حمله قلبی داشته باشد. اگر p-value بیشتر از ۰.۰۵ باشد، نشان‌دهنده عدم وجود تفاوت معنی‌دار در میزان فعالیت بدنی بین دو گروه است.

با مشاهده خروجی درمیابیم که فعالیت فیزیکی در وقوع حمله قلبی موثر است.

۲-۲-۲ تست پیرسون

تست پیرسون یکی از تست‌های آماری مهم در تحلیل داده‌ها است که برای اندازه‌گیری ارتباط خطی بین دو متغیر مستقل استفاده می‌شود. این تست بر اساس ضریب همبستگی پیرسون محاسبه می‌شود که نمایانگر میزان ارتباط خطی بین دو متغیر است. در اصول آماری، این تست به عنوان "آزمون همبستگی پیرسون" (Pearson Correlation Test) شناخته می‌شود.

ضریب همبستگی پیرسون (Pearson Correlation Coefficient) به عنوان "r" نمایش داده می‌شود و در مقادیر بین -۱ تا ۱ قرار دارد:

- اگر r برابر با ۱ باشد، این نشان‌دهنده یک ارتباط خطی مثبت کامل است.
- اگر r برابر با -۱ باشد، این نشان‌دهنده یک ارتباط خطی منفی کامل است.
- اگر r برابر با ۰ باشد، این نشان‌دهنده عدم وجود ارتباط خطی است.

تست پیرسون از مقدار p-value نیز برای ارزیابی اهمیت ارتباط استفاده می‌کند. اگر مقدار p-value به اندازه کافی کم باشد (معمولاً کمتر از ۰.۰۵)، این نشان می‌دهد که ارتباط میان دو متغیر معنی‌دار است و نتیجه آزمون قابل اطمینان است.

از تست پیرسون به عنوان یک ابزار مفید برای بررسی ارتباطات خطی بین متغیرها در تحلیل داده‌های آماری استفاده می‌شود، اما برای ارتباطات غیرخطی به تست‌های دیگر مانند تست همبستگی اسپیرمن و کندال تجزیه و تحلیل می‌آید.

برای نمونه تست پیرسون را برای بررسی اثر فعالیت بدنی بر حمله قلبی انجام میدهیم:

Pearson Test

based on **Activity**

```
from scipy.stats import pearsonr

physical_activity = df['PhysActivity']
heart_disease = df['HeartDiseaseorAttack']

correlation_coefficient, p_value = pearsonr(physical_activity, heart_disease)


print("r : {:.2f}".format(correlation_coefficient))
print("p-value : {:.5f}".format(p_value))
```

```
r : -0.09
p-value : 0.00000
```

شکل۲-۹ نتیجه تست پیرسون براساس فعالیت بدنی

خروجی را تحلیل میکنیم:

۱. r: -0.09

- این بخش از خروجی نمایانگر ضریب همبستگی پیرسون است.

- مقدار منفی نشان می‌دهد که ارتباط میان دو متغیر معکوس است (یعنی افزایش یکی از متغیرها با کاهش دیگری همراه است).

- مقدار ۰.۰۹ نشان‌دهنده شدت این ارتباط خطی است. به عبارت دیگر، این ارتباط نسبتاً ضعیف است.

۲. p-value: 0.00000

- این بخش از خروجی نمایانگر مقدار p-value است که برای ارزیابی اهمیت ارتباط استفاده می‌شود.

- مقدار ۰.۰۰۰۰۰ نشان می‌دهد که مقدار p-value بسیار کم است.

- وقتی مقدار p-value بسیار کم باشد (معمولاً کمتر از ۰.۰۵)، نشان‌دهنده این است که ارتباط میان دو متغیر معنی‌دار است و نتیجه تست قابل اطمینان است.

بنابراین، در این مورد:

- مقدار r منفی و نزدیک به صفر نشان‌دهنده یک ارتباط خطی معکوس و ضعیف بین دو متغیر "میـزان فعالیت بدنی" و "حمله قلبی" است.

- مقدار بسیار کم p-value نشان می‌دهد که این ارتباط معنی‌دار است و احتمـال وقـوع آن بـه صـورت تصادفی نیست.

در اینجا، ارتباط معنی‌دار معکوسی بین میزان فعالیت بدنی و وقوع حمله قلبی وجود دارد، اما این ارتباط بسیار ضعیف است.

برای بررسی حالت دیگر همبستگی کلسترول بالا و حمله قلبی را بررسی میکنیم:

Based on **HighChol**

```
from scipy.stats import pearsonr


HighChol = df['HighChol']
heart_disease = df['HeartDiseaseorAttack']

correlation_coefficient, p_value = pearsonr(HighChol, heart_disease)

print("r : {:.2f}".format(correlation_coefficient))
print("p-value : {:.5f}".format(p_value))
```

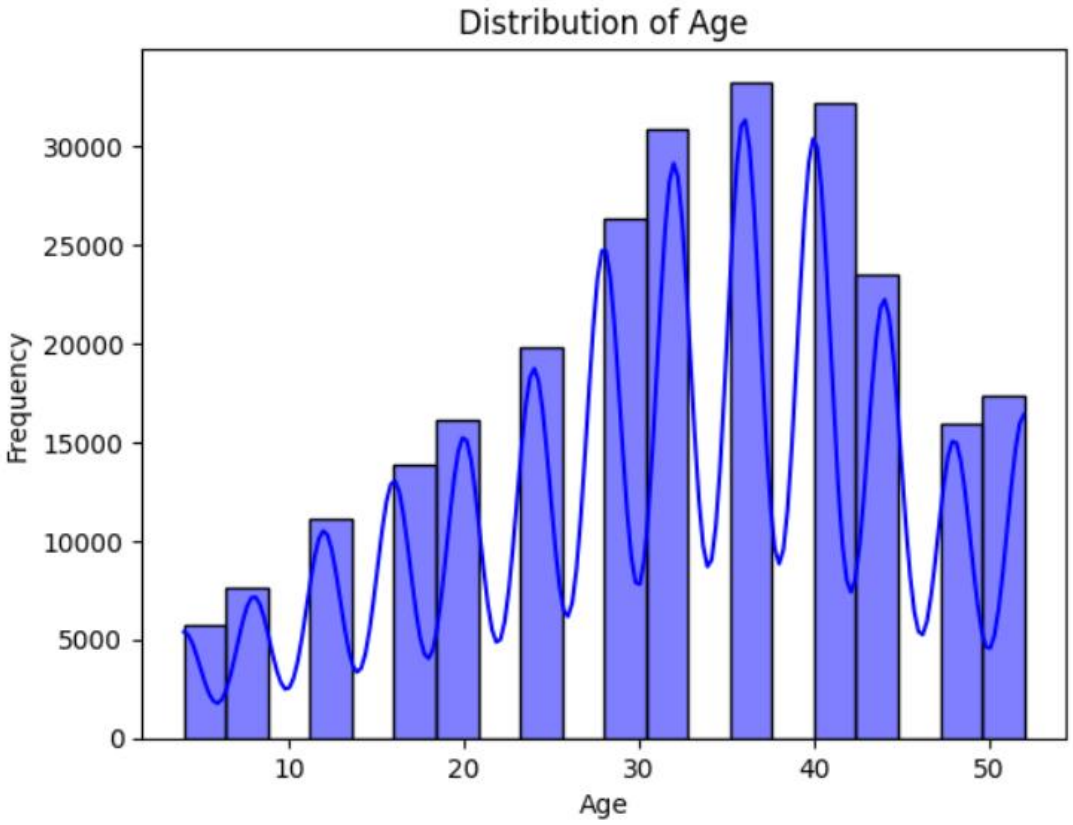
```
r : 0.18
p-value : 0.00000
```

شکل۲-۱۰ نتیجه تست پیرسون براساس کلسترول بالا

نتیجه نشان می‌دهد که کلسترول بالا می‌تواند با افزایش احتمال وقوع حمله قلبی همراه باشد، اگرچه این ارتباط به عنوان یک ارتباط ضعیف تشخیص داده شده است.

فصل سوم: مصورسازی

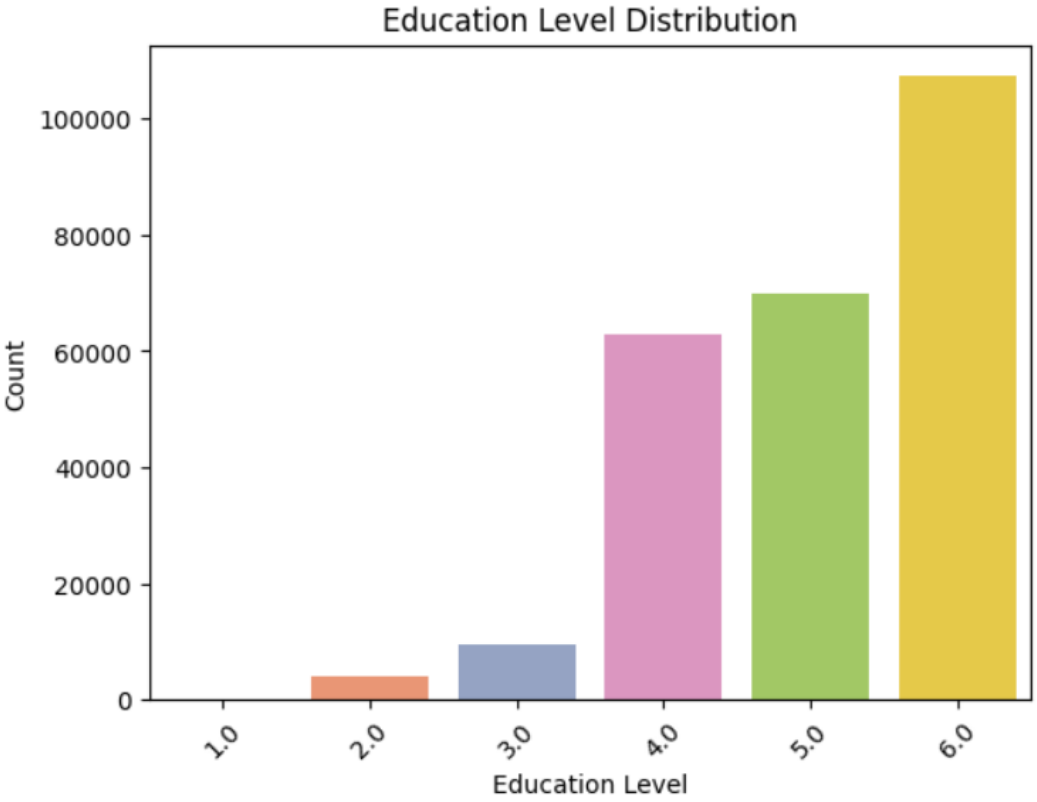
۳-۱- سن



شکل۳-۱ مصورسازی براساس سن

همانطور که مشاهده میشود افراد ۳۰-۴۰ ساله بیشترین احتمال وقوع قلبی را دارند.

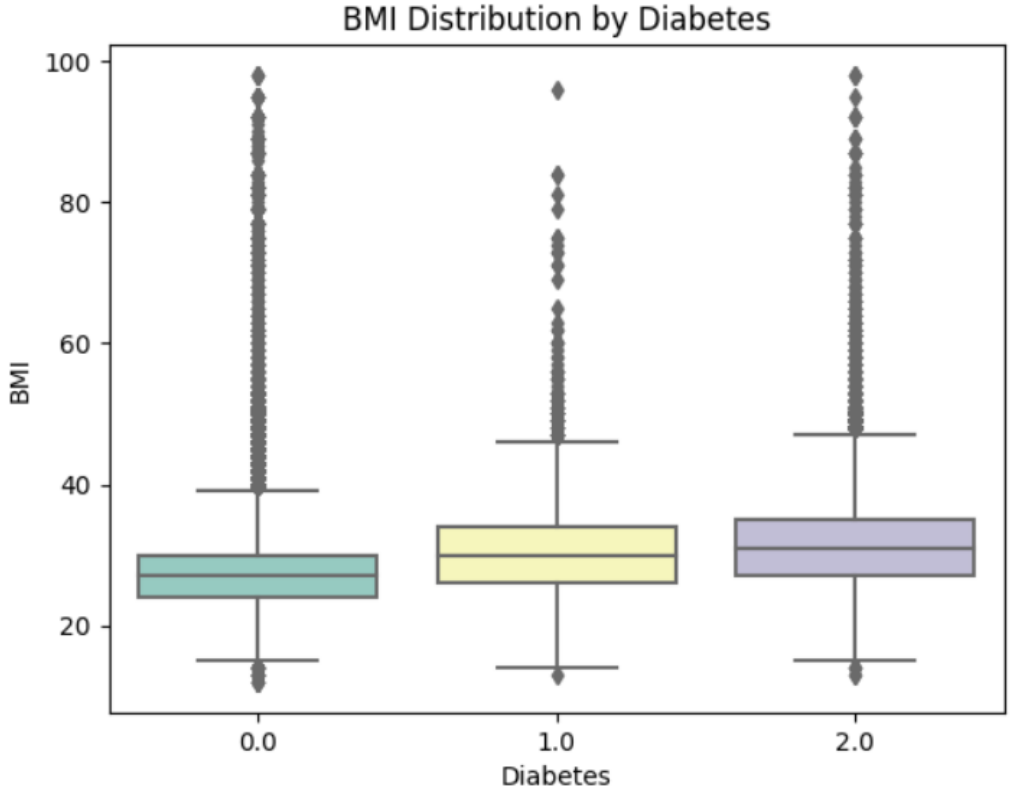
۳-۲- تحصیلات



شکل۳-۲ مصورسازی براساس تحصیلات

همانطور که مشاهده میشود افراد با تحصیلات بالاتر بیشتر مورد حمله قلبی قرار گرفته اند.

۳-۳- BMI براساس دیابت



شکل۳-۳ مصورسازی BMI براساس دیابت

۳-۴- ماتریس همبستگی

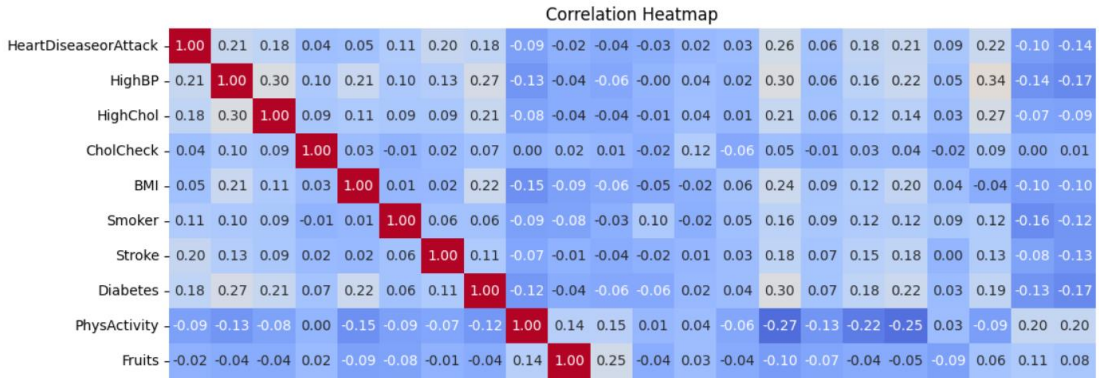
ماتریس همبستگی (Correlation Matrix) یک جدول یا ماتریس آماری است که نمایانگر ارتباط‌ها و همبستگی بین متغیرها در یک دیتاست است

در ماتریس همبستگی، هر خانه نشان‌دهنده مقدار همبستگی بین دو متغیر است. ارتباط بین دو متغیر می‌تواند مثبت یا منفی باشد:

- اگر مقدار همبستگی مثبت باشد (بین ۰ تا +۱)، این نشان‌دهنده یک ارتباط مثبت است؛ یعنی هنگامی که یک متغیر افزایش می‌یابد، متغیر دیگر هم افزایش می‌یابد.

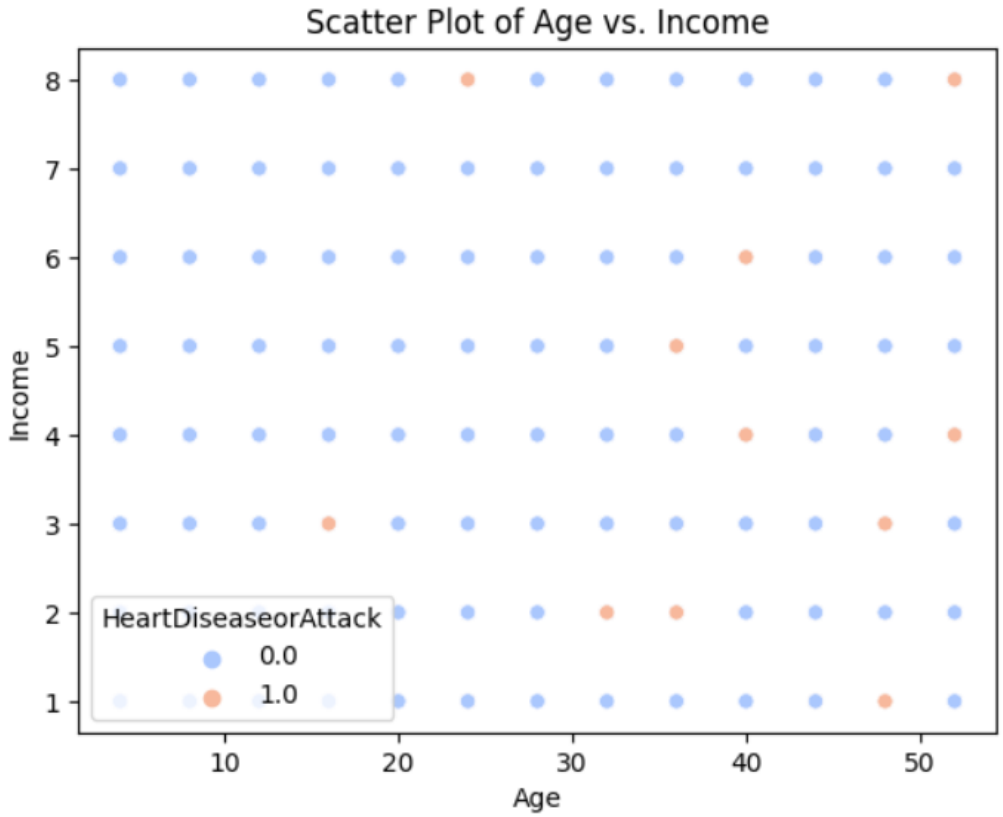
- اگر مقدار همبستگی منفی باشد (بین -۱ تا ۰)، این نشان‌دهنده یک ارتباط منفی است؛ یعنی هنگامی که یک متغیر افزایش می‌یابد، متغیر دیگر کاهش می‌یابد.

- اگر مقدار همبستگی صفر باشد (نزدیک به ۰)، این نشان‌دهنده نبود همبستگی یا ارتباط ضعیف بین دو متغیر است.



شکل۳-۳ بخشی از ماتریس همبستگی

۳-۵- درآمد بر اساس سن



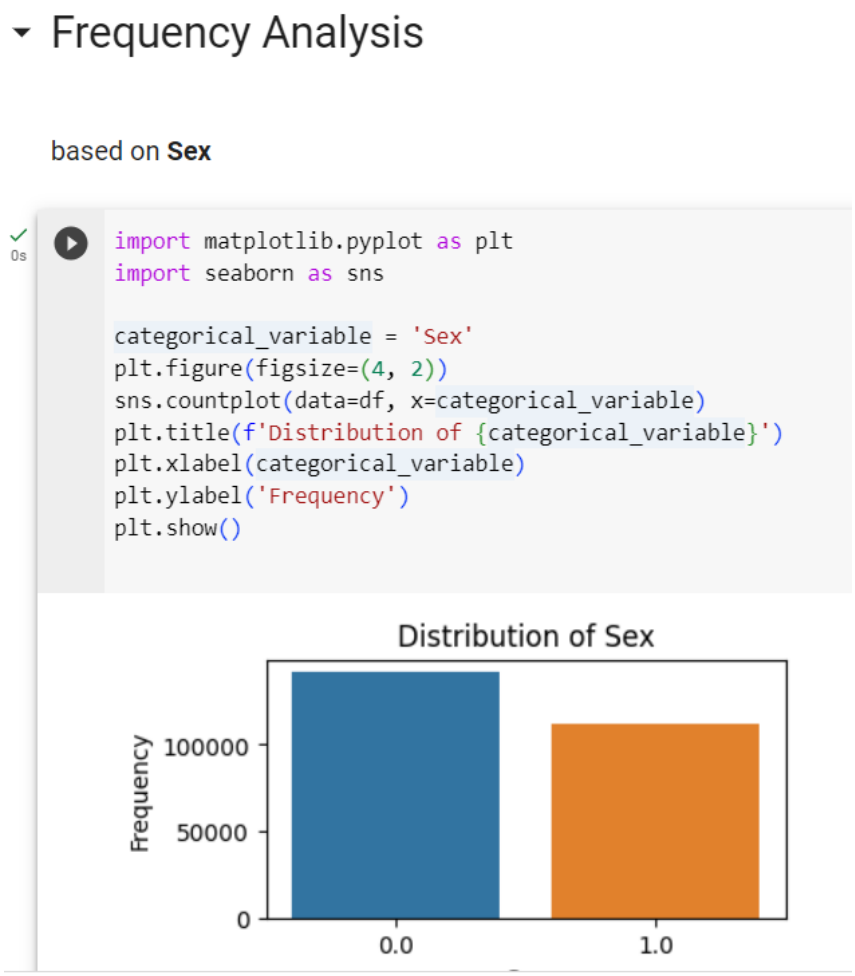
شکل۳-۴ نمودار چگونگی حمله قلبی براسا سن و درآمد

فصل چهارم: تحلیل‌ها

۴-۱-تحلیل توزیع فراوانی

تحلیل توزیع فراوانی یک روش آماری است که برای مطالعه توزیع متغیرهای دسته‌ای (متغیرهایی که مقادیر محدود و مجزایی دارند) استفاده می‌شود. این تحلیل به ما امکان می‌دهد بفهمیم که هر مقدار ممکن از متغیر چندین بار در داده‌ها ظاهر می‌شود و نسبت هر مقدار به کل داده‌ها چقدر است.

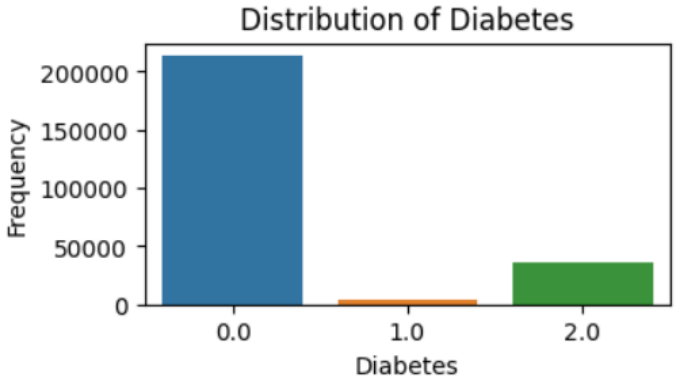
به تحلیل توزیع فراوانی براساس چند ویژگی میپردازیم:



شکل۴-۱ تحلیل توزیع فراوانی بر اساس جنسیت

based on **Diabetes**

```
[40] categorical_variable = 'Diabetes'
     plt.figure(figsize=(4, 2))
     sns.countplot(data=df, x=categorical_variable)
     plt.title(f'Distribution of {categorical_variable}')
     plt.xlabel(categorical_variable)
     plt.ylabel('Frequency')
     plt.show()
```



شکل۴-۲ تحلیل توزیع فراوانی بر اساس دیابت

۴-۲-تحلیل رگرسیون

این کد به مدل رگرسیون لجستیک (Logistic Regression) می‌پردازد و از آن برای پیش‌بینی بیماری قلبی یا حمله قلبی استفاده می‌کند. در اینجا توضیح کارهایی که این کد انجام می‌دهد آمده است:

۱. انتخاب متغیرهای مستقل:

- متغیرهای مستقل از دیتاست انتخاب می‌شوند. این متغیرها به عنوان ویژگی‌های ورودی به مدل برای پیش‌بینی بیماری قلبی مورد استفاده قرار می‌گیرند.

۲. تقسیم داده به دو بخش آموزش و آزمون:

- داده‌ها به دو بخش تقسیم می‌شوند: بخش آموزش برای آموزش مدل و بخش آزمون برای ارزیابی عملکرد مدل.

۳. ایجاد مدل رگرسیون لجستیک:

- یک مدل رگرسیون لجستیک ایجاد می‌شود. این مدل اجازه می‌دهد تا احتمال حضور یا عدم حضور بیماری قلبی را بر اساس ویژگی‌های ورودی محاسبه کرد.

۴. آموزش مدل:

- مدل روی داده‌های آموزش آموزش داده می‌شود تا بیاموزد چگونه ویژگی‌های ورودی را با احتمال حضور یا عدم حضور بیماری قلبی مرتبط کند.

۵. پیش‌بینی با استفاده از مدل:

- مدل بر روی داده‌های آزمون پیش‌بینی انجام می‌دهد و احتمال حضور یا عدم حضور بیماری قلبی برای هر نمونه آزمون محاسبه می‌شود.

۶. ارزیابی مدل:

- عملکرد مدل ارزیابی می‌شود. در اینجا، معیارهایی مانند دقت (Accuracy)، ماتریس اشتباهات (Confusion Matrix) و گزارش طبقه‌بندی (Classification Report) برای ارزیابی عملکرد مدل استفاده می‌شوند. این معیارها به شما اطلاعاتی ارائه می‌دهند که درک بهتری از عملکرد مدل رگرسیون لجستیک روی داده‌های آزمون را فراهم می‌کنند.

Classification Report:

	precision	recall	f1-score	support
0.0	0.92	0.99	0.95	45968
1.0	0.51	0.13	0.21	4768
accuracy			0.91	50736
macro avg	0.71	0.56	0.58	50736
weighted avg	0.88	0.91	0.88	50736

شکل۴-۳ تحلیل رگرسیون

گزارش طبقه‌بندی (Classification Report) نتایج عملکرد مدل رگرسیون لجستیک را برای هر کلاس (بیماری قلبی یا عدم حضور بیماری قلبی) نمایش می‌دهد. این گزارش شامل معیارهایی نظیر

دقت (Precision)، بازیابی (Recall) و اسکور F1 (F1-Score) برای هر کلاس و همچنین دقت کلی (Accuracy) مدل است.

- Precision (دقت): نسبت تعداد نمونه‌هایی که به درستی تشخیص داده شده‌اند (True Positive) به تعداد نمونه‌هایی که به عنوان مثبت تشخیص داده شده‌اند (True Positive + False Positive) است. برای کلاس ۰.۰ دقت ۰.۹۲ و برای کلاس ۱.۰ دقت ۰.۵۱ است. این معیار نشان‌دهنده توانایی مدل در تشخیص نمونه‌های واقعی کلاس را نشان می‌دهد.

- Recall (بازیابی): نسبت تعداد نمونه‌هایی که به درستی تشخیص داده شده‌اند (True Positive) به تعداد نمونه‌هایی که به درستی در کلاس مورد نظر هستند (True Positive + False Negative) است. برای کلاس ۰.۰ بازیابی ۰.۹۹ و برای کلاس ۱.۰ بازیابی ۰.۱۳ است. این معیار نشان‌دهنده توانایی مدل در شناسایی تمام نمونه‌های مثبت واقعی کلاس را نشان می‌دهد.

- F1-Score: اسکور F1 یک ترکیب از دقت و بازیابی است و به صورت زیر محاسبه می‌شود: ۲ * (Precision * Recall) / (Precision + Recall). برای کلاس ۰.۰ اسکور F1 برابر با ۰.۹۵ و برای کلاس ۱.۰ اسکور F1 برابر با ۰.۲۱ است. این معیار نشان‌دهنده تعادل بین دقت و بازیابی مدل است.

- Accuracy (دقت کلی): نسبت تعداد نمونه‌هایی که به درستی تشخیص داده شده‌اند (True Positive + True Negative) به تعداد کل نمونه‌ها (تمام موارد) است. دقت کلی مدل ۰.۹۱ است و نشان‌دهنده توانایی مدل در تشخیص صحیح کل نمونه‌هاست.

- Support (تعداد نمونه‌ها): تعداد نمونه‌هایی که به هر کلاس تعلق دارند نمایش داده می‌شود.

به طور خلاصه، گزارش طبقه‌بندی به شما اطلاعاتی ارائه می‌دهد که می‌توانید برای ارزیابی عملکرد مدل در تفکیک بیماری قلبی و عدم حضور آن بهره‌برد. به عنوان مثال، دقت برای کلاس ۰.۰ بسیار بالاست که نشان‌دهنده توانایی مدل در تشخیص صحیح افراد بدون بیماری قلبی است، اما بازیابی برای کلاس ۱.۰ کم است که نشان‌دهنده نقاط ضعف در تشخیص افراد دارای بیماری قلبی است.

فصل پنجم

جمع‌بندی و نتیجه‌گیری و پیشنهادات

جمع‌بندی و نتیجه‌گیری

با توجه به تحلیل دیتاست بیماری قلبی که انجام دادیم، می‌توانیم به نتیجه‌گیری‌های زیر برسیم:

۱. تحلیل آماری داده:

- دیتاست شامل اطلاعاتی از افراد در مورد وضعیت سلامتی و عوامل مختلفی مثل کلسترول، BMI، عادت‌های سیگاری، و میزان فعالیت بدنی است.

۲. تحلیل میانگین‌ها با ازمون تی:

- میانگین مقادیر برای ویژگی‌ها مورد مطالعه قرار گرفت.

- می‌توان نتیجه گرفت که برای بسیاری از ویژگی‌ها مقادیر میانگین نسبت به حمله قلبی و عدم حمله قلبی تفاوت دارند.

۴. آزمون همبستگی پیرسون:

- آزمون همبستگی پیرسون نشان داد که وجود کلسترول بالا (HighChol) با حمله قلبی مرتبط است، اما این ارتباط ضعیف است و فعالیت ورزشی رابطه عکس با حمله قلبی دارد.

۵. تحلیل توزیع فراوانی:

- چگونگی توزیع ویژگی ها براساس وقوع حمله قلبی را مورد بررسی داد.

۶. مدل‌سازی و آزمون رگرسیون لجستیک:

- مدل رگرسیون لجستیک برای پیش‌بینی حمله قلبی براساس ویژگی‌های مختلف تشکیل شد.

- مدل با دقت خوبی بیماری قلبی و عدم حضور آن را تشخیص داد. با این حال، در تشخیص افراد دارای بیماری قلبی، عملکرد مدل کمتر بود.

با توجه به این نتایج، می‌توانیم نقاط ضعف و قوت مدل را شناسایی کرده و به تصمیم‌گیری‌های مرتبط با بیماری قلبی کمک کنیم. این تحلیل‌ها می‌توانند به پزشکان و محققان در تفسیر داده‌ها و ارتقاء تشخیص و پیش‌بینی بیماری‌های قلبی کمک کنند.

منابع و مراجع

[۱] https://www.kaggle.com/code/aniketkadam702030/heart-disease-prediction-90-accuracy

[2] https://www.kaggle.com/code/precioux/heart-disease-health-indicators

[3] https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset

پیوست‌ها

- دیتاست

https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset

- کد

https://colab.research.google.com/drive/1lHbFHnPN06vezPMcwqf8rxz_AOp_zAGI?usp=sharing

Abstract

In this project, we want to analyze a data set of people with biological characteristics to classify the occurrence of heart attack. Our main goal in this project will be the statistical analysis of the data, its visualization and analysis.

Keywords:

Visualization, heart disease, data mining, data analysis

Amirkabir University of Technology
(Tehran Polytechnic)

Department of Computer Science

Project 2

A Study in Heart Disease Health Indicators Dataset

By
Samin Mahdipour

Supervisor
Dr.Ghatee

October 2023