فصل دوم
یادگیری مدل یا مدل‌سازی

آرزو حبیبی راد

گروه آمار

دانشگاه فردوسی مشهد

- ✓ بعد از آماده سازی و مصور سازی داده ها، هدف بعدی طراحی مدل‌های مناسب برای پیش بینی می باشد.
- ✓ روشهای مدل‌سازی به دو دسته اصلی و **« یادگیری با ناظر»** و **«یادگیری بدون ناظر»** تقسیم می شوند.

- Supervised learning
- Un supervised learning

یادگیری با ناظر:

- متغیر یا ستونی که مشخص کند رکورد مورد نظر به کدام دسته یا کلاس تعلق دارد را متغیر برچسب می گویند. برای برچسب دادن به رکوردها نیاز به ناظر است.
- اگر در پایگاه داده ها، متغیر (یا فیلد) بر چسب‌دار وجود داشته باشد برای مدل‌سازی این نوع از پایگاه داده‌ها می توان از روش‌های یادگیری با ناظر استفاده کرد.

معروف‌ترین روش های یادگیری با ناظر:

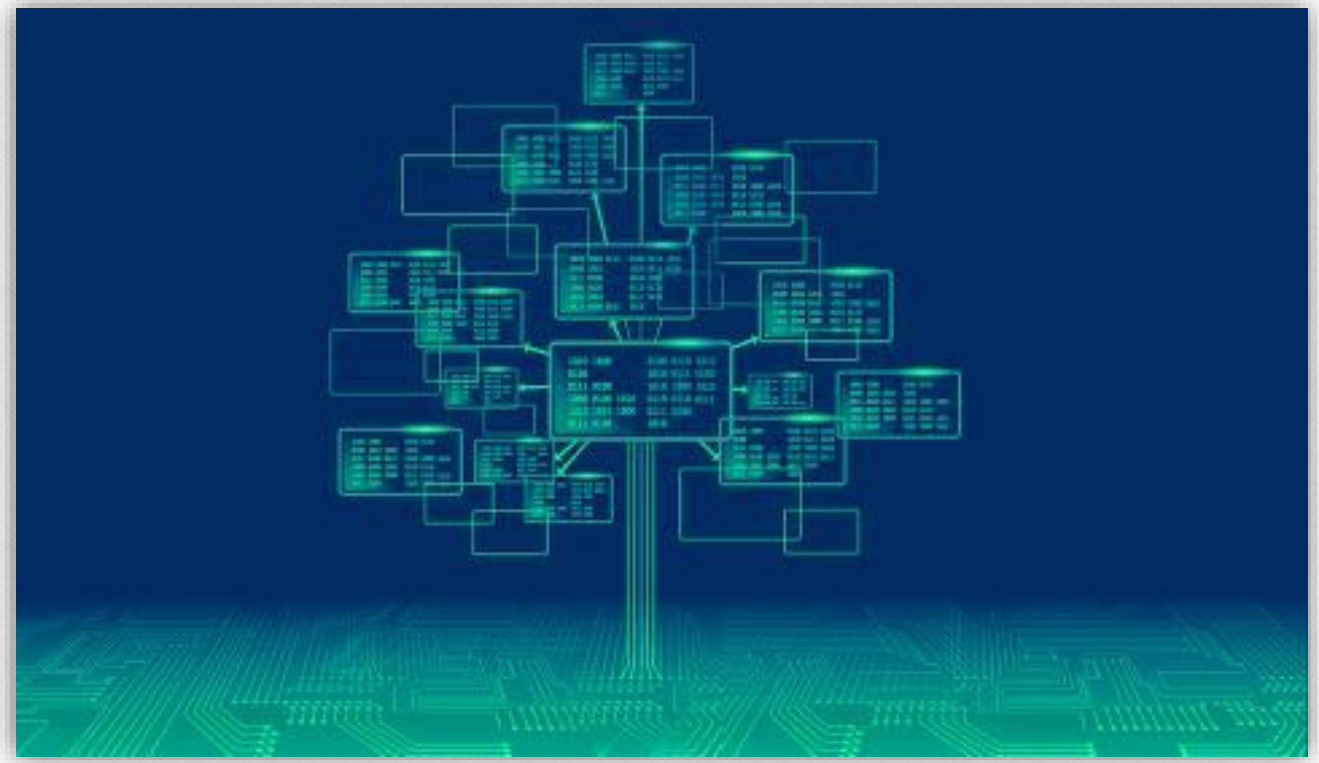
1. دسته بندی
2. رگرسیون لوژستیک دو سطحی
3. رگرسیون لوژستیک چند سطحی

دسته بندی:

- ✓ در دسته بندی برای هرکدام از رکوردهای مجموعه داده مورد کاوش، یک برچسب که بیان‌گر حقیقتی است، وجود دارد.
- ✓ هدف الگوریتم یادگیری پیدا کردن نظم حاکم بر انواع برچسب‌ها براساس سایر ویژگی‌های رکورد می‌باشد.
- ✓ الگوریتم‌های دسته بندی شامل دو مرحله **آموزش** و **ارزیابی** هستند.

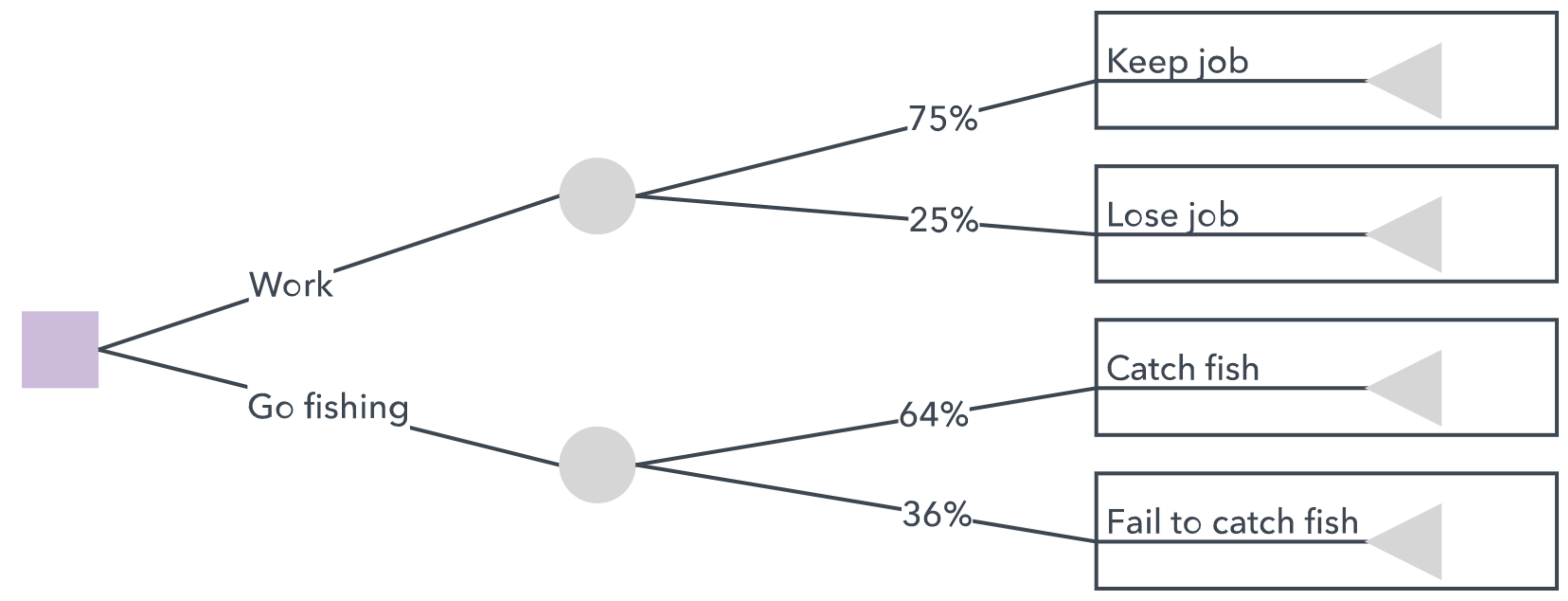
- در مرحله آموزش براساس داده‌های آموزشی یک مدل ساخته می‌شود و در مرحله آزمایش براستفاده از داده‌های آزمایشی، **دقت** و **کارایی** مدل ساخته شده مورد ارزیابی قرار خواهد گرفت.
- در این فصل مرحله اول یعنی ساخت مدل مورد بحث قرار می‌گیرد و مرحله دوم در فصل آینده بررسی می‌شود.

دسته بندی با کمک درخت تصمیم

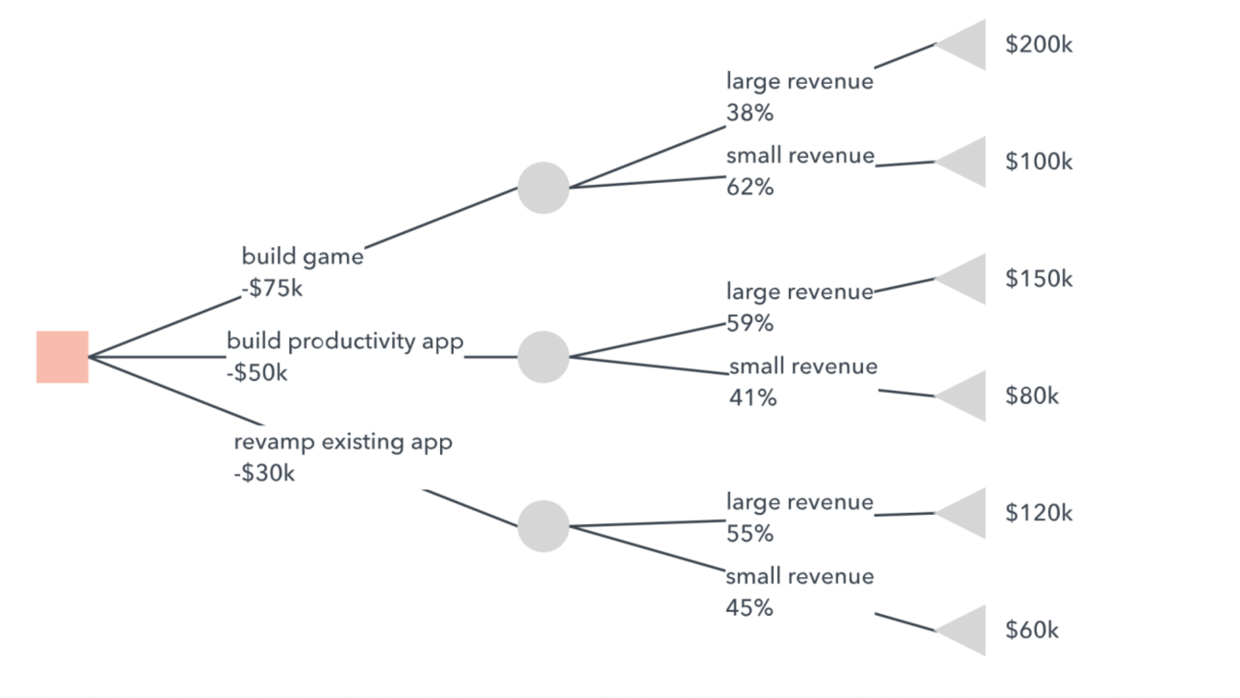


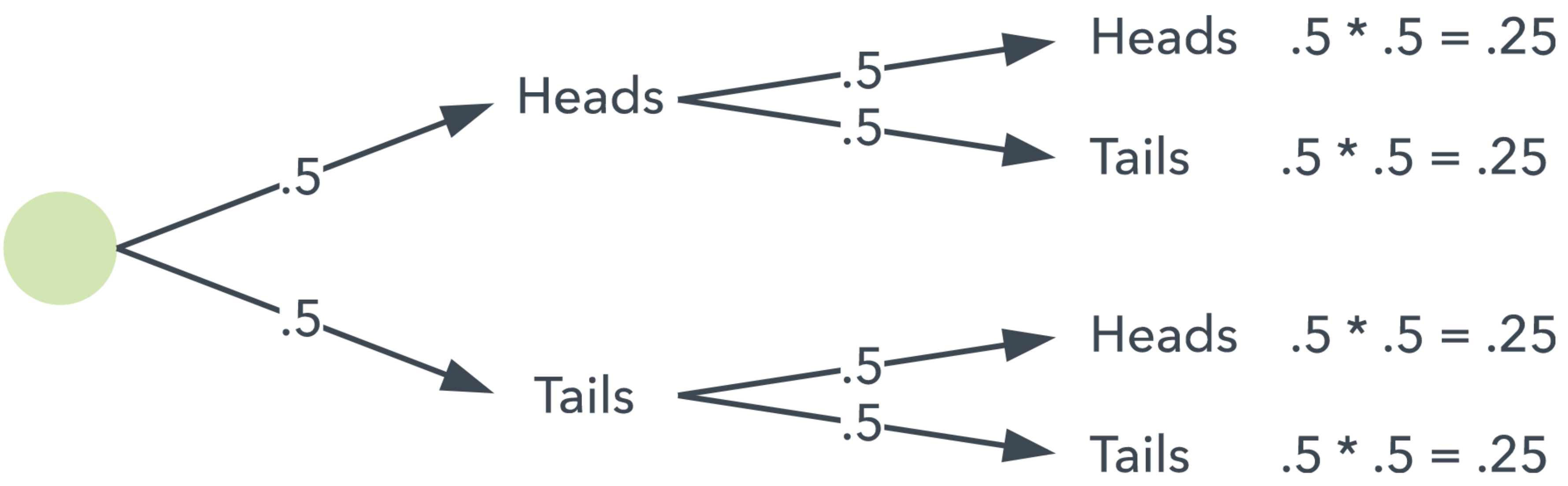
چرا درخت تصمیم:

- ✓ یکی از روش‌های یادگیری با ناظر برای تحلیل داده‌ها، روش دسته بندی داده‌ها می باشد.
- ✓ درخت تصمیم، یکی از پر کاربردترین روش‌های یادگیری با ناظر در قسمت دسته بندی داده‌ها می باشد به طوری که ساختار درختی شبیه فلوچارت دارد.
- ✓ این درخت قادر به تولید توصیفاتی قابل درک از روابط موجود در یک مجموعه داده است.
- ✓ از این روش می توان برای دسته بندی و پیش بینی استفاده کرد.
- ✓ این درخت برای یک مجموعه داده همواره یکتا نیست.



Work
Go fishing
75%
25%
64%
36%
Keep job
Lose job
Catch fish
Fail to catch fish


$200k
large revenue
38%
small revenue
62%
$100k
build game
-$75k
$150k
large revenue
59%
build productivity app
-$50k
small revenue
41%
$80k
revamp existing app
-$30k
large revenue
55%
$120k
small revenue
45%
$60k


Heads
Tails
.5
.5
.5
.5
.5
.5
Heads .5 * .5 = .25
Tails .5 * .5 = .25
Heads .5 * .5 = .25
Tails .5 * .5 = .25

رسم درخت تصمیم:

- ✓ درخت تصمیم یک روش یادگیری **با ناظر** و از روشهای قدیمی و معروف برای ساخت مدل دسته بندی است و بسیاری از الگوریتم های طبقه بندی بر پایه ی این درخت‌ها ساخته شده اند.
- ✓ **Decision Tree** مفهومی است که اگر در نظر دارید تا تصمیم پیچیده ای بگیرید و یا می‌خواهید مسائل را برای خودتان به بخشهای کوچک تری تقسیم کرده تا به شکل بهتری قادر به حل آن ها گردیده و ذهن تان را سازمان‌دهی کنید، می‌توانید از آن استفاده نمایید.
- ✓ یادگیری درخت تصمیم روشی برای **تقریب توابع هدف با مقادیر گسسته** است. این روش نسبت به نویز داده‌ها مقاوم بوده و قادر است ترکیب فصلی گزاره های عطفی را یاد بگیرد. درخت تصمیم درختی است که در آن نمونه ها را به نحوی دسته بندی میکند که از ریشه به سمت پائین رشد میکنند و در نهایت به گره های برگ میرسد.

یک درخت تصمیم شامل سه نوع گره است:

1) **گره ریشه:** گره‌ای که هیچ یالی به آن وارد نمی‌شود و ممکن است تعدادی یال از آن خارج شوند.
2) **گره داخلی:** دقیقا یک یال به آن وارد شده و تعداد دو یا بیشتر یال از آن خارج می‌شود.
3) **گره برگ:** دقیقا یک یال به آن وارد شده و هیچ یالی از آن خارج نمی‌شود.

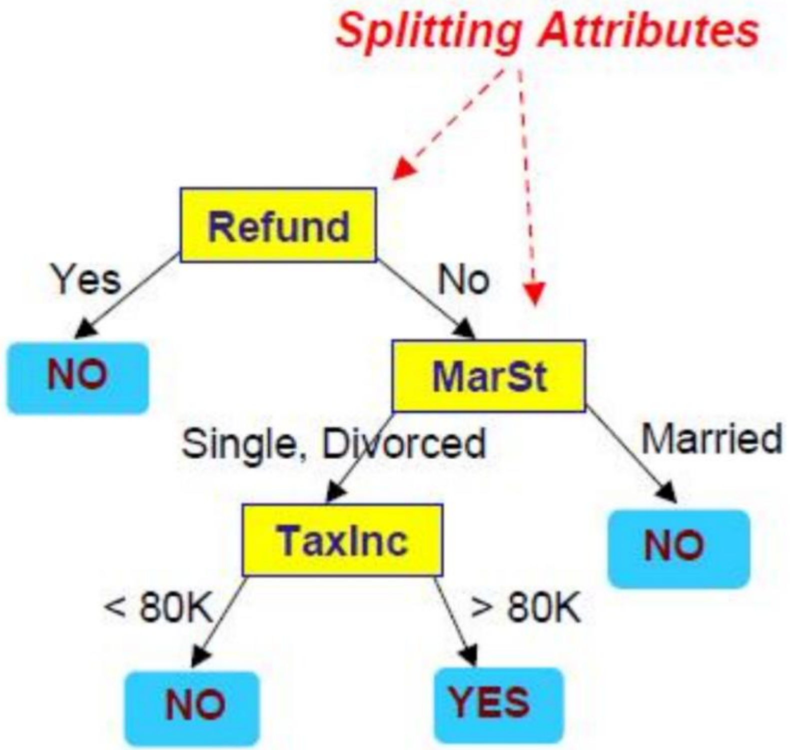
- علت نامگذاری این روش با درخت تصمیم این است که این درخت فرآیند تصمیم گیری برای تعیین دسته یک مثال ورودی را نشان می‌دهد.

- فرض کنید مدیر یک پایگاه هستید و می‌خواهید یک طبقه بندی انجام دهید که بر اساس آن تشخیص دهید که آیا فرد، فرار مالیاتی داشته است یا خیر؟
- این طبقه بندی به ما کمک می‌کند که جلوی فرار مالیاتی را بگیریم . بنابراین می‌خواهیم یک مدل را بر اساس مجموعه داده موجود آموزش دهیم .
- در شکل آمده در اسلاید بعد، مجموعه داده آموزش و یک نمونه درخت تصمیم از روی داده ها نشان داده شده است.

categorical categorical continuous class

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



Model: Decision Tree

الگوریتم ID 3 در رسم درخت تصمیم:

- ✓ **الگوریتم ID 3** یکی از الگوریتم های پایه ای و ساده برای ساختِ درخت‌های تصمیم است.
- ✓ در یک درخت تصمیم، مهم است که کدام یک از ویژگی‌ها (یا همان متغیر) را در سطوح بالاتری از درخت انتخاب کنیم تا به طبقه بندی کمک کند.
- برای رسم درخت تصمیم لازم است بدانیم کدام متغیر اطلاعات بیشتری در مورد متغیر برچسب می دهد تا در سطح بالاتری قرار گیرد. این امر با کمک مفاهیم متفاوتی صورت می گیرد از جمله بهره اطلاع (**Information Gain**).
- متغیری که داری بیشترین بهره اطلاعاتی باشد در گره ریشه درخت تصمیم قرار می‌گیرد و این روند تا رسیدن به برگ (سطوح متغیر برچسب) ادامه می‌یابد.
- **الگوریتم ID 3** برای متغیر های برچسبی که پیوسته باشند ساخته نشده است.
- برای محاسبه بهره اطلاعاتی به مفاهیم در نظریه اطلاع از جمله آنتروپی و جینی نیاز داریم.

محاسبه بهره اطلاعاتی:

- روش های متفاوتی برای محاسبه Information Gain وجود دارد از جمله استفاده از مفاهیم آنتروپی و جینی.
- فرض کنید متغیر برچسب (L) با m سطح و برای متغیر دلخواه (A) با سطح داریم:
- بهره اطلاعاتی متغیر A

$$\text{Information Gain(A)}=\text{Ent/Gini(L)}-\text{Ent/Gini(A)}$$

- هرچه بهره اطلاعاتی متغیری بیشتر باشد، متغیر در سطح بالاتری در درخت تصمیم قرار می گیرد.
- به عبارتی، هرچه میزان انتروپی (جینی) متغیر A کمتر باشد، بهره اطلاعاتی آن بیشتر می شود.

- آنتروپی متغیر گسسته برچسب **(L)** با m سطح و n رکورد:

$$\boldsymbol{En(L) = -\sum_{j=1}^{m} p_j \ln(p_j)}$$

- که در آن $\boldsymbol{p_j}$ احتمال سطح زام متغیر برچسب می باشد.
- آنتروپی متغیر دلخواه **(A)** با k سطح:

$$\boldsymbol{En(A) = \sum_{i=1}^{k} \frac{n_i}{n} En(A_i)}$$

- که در آن آنتروپی هر سطح متغیر دلخواه $\boldsymbol{(A_i)}$ برابربا:

$$\boldsymbol{En(A_i) = -\sum_{j=1}^{m} p_j(A_i) \ln p_j(A_i)}$$

و در آن $\boldsymbol{p_j(A_i)}$ احتمال سطح زام متغیر $\boldsymbol{A}$ در هریک از سطوح متغیر برچسب زام می باشد.

مثال: 14 رکورد زیر مربوط به خرید کامپیوتر برای متغیرهای اعتبار بانکی، سن، درآمد و تحصیل مشتریان

14 : خرید | خیر

	اعتبار	سن	درآمد	تحصیل	خرید	var
1	خوب	جوان	بالا	خوب	خیر	
2	عالی	جوان	بالا	خوب	خیر	
3	خوب	میانسال	بالا	خوب	بله	
4	خوب	مسن	متوسط	خوب	بله	
5	خوب	مسن	بالا	بد	بله	
6	عالی	مسن	بالا	بد	خیر	
7	عالی	میانسال	بالا	بد	بله	
8	خوب	جوان	متوسط	خوب	خیر	
9	خوب	جوان	بالا	بد	بله	
10	خوب	مسن	متوسط	بد	بله	
11	عالی	جوان	متوسط	بد	بله	
12	عالی	میانسال	متوسط	خوب	بله	
13	خوب	میانسال	بالا	بد	بله	
14	عالی	مسن	متوسط	خوب	خیر	
15						
16						

بله=(+)   خیر=(-) : خرید $L = \{(-)5$ و $(+)9\}$

En(L)=-{5/14ln(5/14)+9/14ln(9/14)}=0.652

اعتبارمالی : $A = \{$خوب 8, عالی 6$\} \rightarrow$ $A1: \{(-)2$و$(+)6\}$ = خوب $A2: \{(-)3(+)$و$3\}$ = عالی

سن : $B = \{$مسن 5, میانسال 4, جوان 5$\} \rightarrow$ $B1: \{(+)4$و$(-)0\}$ = میانسال ، $B2: \{(+)3$و$(-)2\}$ = مسن ، $B3: \{(-)3$و$(+)\ 2\}$ = جوان

درآمد : $C = \{$متوسط 6, بالا 8$\} \rightarrow$ $C1: \{(+)4$ و $(-)2\}$ = متوسط ، $C2: \{(+)5$ و $(-)3\}$ = بالا

تحصیل : $D = \{$خوب 7 و بد 7$\} \rightarrow$ $D1: \{(+)6$ و $(-)\ 1\}$ = بد ، $D2: \{(+)3$ و $(-)4\}$ = خوب

En(A1)=-{2/8ln(2/8)+6/8ln(6/8)}=-{-0.346-0.216}=0.562
En(A2)=-{3/6ln(3/6)+3/6ln(3/6)}=-{-0.346-0.346}=0.693
Entropy(A)={8/14(.562)+6/14(.693)}=0.321+0.297=0.618
Gain(A)=En(L)-En(A)=.652-.618=0.007

En(B1)=-{0+4/4ln(4/4)}=0
En(B2)=-{3/5ln(3/5)+2/5ln(2/5)}=0.673
En(B3)=-{3/5ln(3/5)+2/5ln(2/5)}=0.673
Entropy(B)={4/14(0)+5/14(.673)+5/14(.673)}=0.481
Gain(B)=En(L)-En(B)=.652-.481=.171

$$En(C_1) = -\left\{\frac{4}{6}\ln\left(\frac{4}{6}\right) + \frac{2}{6}\ln\left(\frac{2}{6}\right)\right\} = 0.637$$

$$En(C_2) = -\left\{\frac{5}{8}\ln\left(\frac{5}{8}\right) + \frac{3}{8}\ln\left(\frac{3}{8}\right)\right\} = 0.662$$

$$Entropy(C) = \left\{\frac{6}{14}(0.637) + \frac{8}{14}(0.662)\right\} = 0.651$$

$$Gain(C) = En(L) - En(C) = 0.652 - 0.651 = 0.001$$

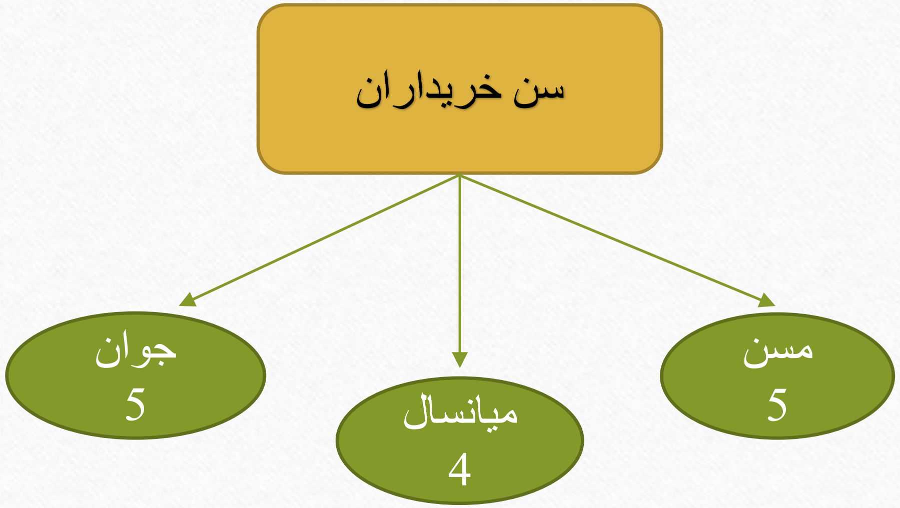
$$En(D_1) = -\left\{\frac{6}{7}\ln\left(\frac{6}{7}\right) + \frac{1}{7}\ln\left(\frac{1}{7}\right)\right\} = 0.41$$

$$En(D_2) = -\left\{\frac{4}{7}\ln\left(\frac{4}{7}\right) + \frac{3}{7}\ln\left(\frac{3}{7}\right)\right\} = 0.683$$

$$Entropy(D) = \left\{\frac{7}{14}(0.41) + \frac{7}{14}(0.683)\right\} = 0.546$$

$$Gain(D) = En(L) - Entropy(D) = 0.652 - 0.546 = 0.106$$

درنتیجه بیشترین بهره اطلاعاتی در خصوص متغیر بر چسب، مربوط به متغیر سن می باشد، لذا متغیر سن در ریشه درخت تصمیم قرار می گیرد.


سن خریداران
جوان
5
میانسال
4
مسن
5

اشکال روش آنتروپی:

- معیار آنترپی در درخت تصمیم گرایش زیادی به انتخاب متغیرهایی دارد که دارای سطوح بیشتری هستند. این یک ضعف برای معیارآنتروپی است.
- مثال: متغیری مانند کد کالا که تعداد سطوح زیادی دارد، در این روش نسبت به سایر متغیرها دارای بهره اطلاعاتی بیشتر بوده و به نادرستی در سطح بالاتر درخت قرار می گیرد.
- برای رفع این مشکل از **اطلاع شکست (Split Information)** استفاده می کنند.

اطلاع شکست:

- برای رفع مشکل آنتروپی از نوعی روش نرمال ساز بر روی معیار آنتروپی استفاده می‌شود که به آن اطلاع شکست می‌گویند.
- برای متغیر دلخواه **A** با **k** سطح داریم:

$$S.I(A) = -\sum_{i=1}^{k} \frac{n_i}{n} \ln(\frac{n_i}{n})$$

- که همان آنتروپی متغیر **A** می باشد.
- بهره اطلاعاتی برای متغیر **A** در این روش با کمک فرمول زیر محاسبه می‌شود:

$$Gainratio(A) = \frac{Gain(A)}{S.I(A)}$$

معیار جینی:

- ✓ جینی متغیر گسسته برچسب (L) با **m** سطح و **n** رکورد:

$$Gini(L) = 1 - \sum_{j=1}^{m} p_j^2$$

که در آن p_j احتمال سطح **j**ام متفیر برچسب می باشد.

- ✓ آنتروپی متغیر دلخواه (A) با **k** سطح:

$$Gini(A) = \sum_{i=1}^{k} \frac{n_i}{n} Gini(A_i)$$

- ✓ که در آن آنتروپی هر سطح متغیر دلخواه (A_i) برابر:

$$Gini(A_i) = 1 - \sum_{j=1}^{m} p_j^2(A_i)$$

که در آن $p(A_i)$احتمال سطح **i**ام متغیر **A** در هریک از سطوح متغیر برچسب می باشد.

مثال: برای 12 رکورد اطلاعات زیر مفروض است، مطلوبست محاسبه بهره اطلاعاتی متغیر A با کمک معیار جینی؟

	lable	a	b	var
1	.00	a1	b1	
2	.00	a1	b2	
3	.00	a1	b2	
4	.00	a1	b3	
5	.00	a1	b3	
6	.00	a2	b3	
7	1.00	a1	b1	
8	1.00	a1	b1	
9	1.00	a2	b1	
10	1.00	a2	b2	
11	1.00	a2	b3	
12	1.00	a2	b3	
13				
14				

- $Gini(l) = 1 - \left\{ \left(\frac{6}{12}\right)^2 + \left(\frac{6}{12}\right)^2 \right\} = 0.5$
- $Gini(a_1) = 1 - \left\{ \left(\frac{5}{7}\right)^2 + \left(\frac{2}{7}\right)^2 \right\} = 0.41$
- $Gini(a_2) = 1 - \left\{ \left(\frac{1}{5}\right)^2 + \left(\frac{4}{5}\right)^2 \right\} = 0.32$
- $Gini(A) = \frac{7}{12}(0.41) + \frac{5}{12}(0.32) = 0.372$
- $Gain(A) = 0.5 - 0.372 = 0.128$

❖بهره اطلاعاتی متغیر B به عنوان تمرین محاسبه شود؟ همچنین مطلوبست رسم درخت تصمیم ؟

محاسبه معیار جینی برای متغیر پیوسته A

- **مثال:** 10 رکورد با متغیر برچسب تقلب در پرداخت مالیات (به صورت بلی و خیر) ومتغیر پیوسته درآمد در جدول زیر آمده اند. با کمک معیار جینی مقداری از این متغیر که بهره اطلاعاتی بیشتری دارد را مشخص کنید.
- **مالیات:** خیر | خیر | خیر | بلی | بلی | بلی | خیر | خیر | خیر | خیر
- **درآمد:** 60 70 75 85 90 95 100 120 125 220
- **وسط درآمد:** 55 65 72 80 87 92 97 110 122 172 230

<> <>

- تعداد پاسخ بله: 0
- تعداد پاسخ خیر: 0

- $Gini(L) = 1 - \left\{\left(\frac{3}{10}\right)^2 + \left(\frac{7}{10}\right)^2\right\} = 0.42$
- $Gini(55 -) = 1 - \left\{\left(\frac{0}{0}\right)^2 + \left(\frac{0}{0}\right)^2\right\} = a \quad Gini(55 +) = 1 - \left\{\left(\frac{3}{10}\right)^2 + \left(\frac{7}{10}\right)^2\right\} = 0.42$
- $Gini(55) = \frac{0}{10}(a) + \frac{10}{10}(0.42) = 0.42$
- $Gini(65 -) = 1 - \left\{\left(\frac{0}{1}\right)^2 + \left(\frac{1}{1}\right)^2\right\} = 0 \quad Gini(65 +) = 1 - \left\{\left(\frac{3}{9}\right)^2 + \left(\frac{6}{9}\right)^2\right\} = 0.46$
- $Gini(65) = \frac{1}{10}(0) + \frac{9}{10}(0.46) = 0.41$

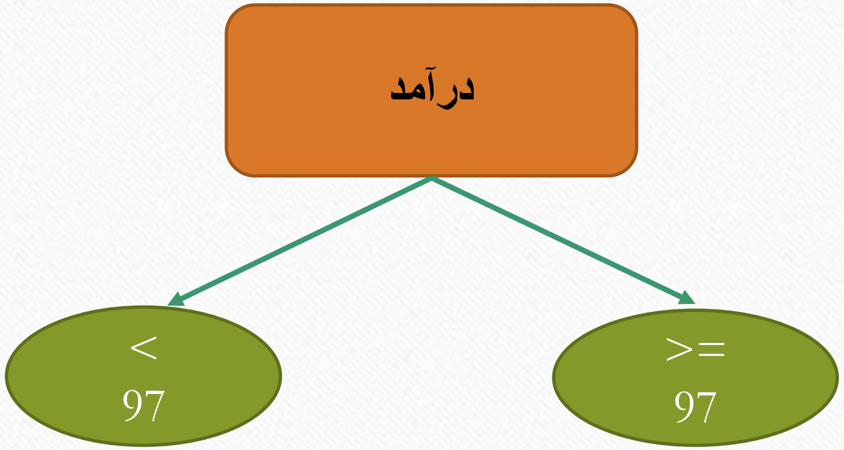
.

.

.

- $Gini(97 -) = 1 - \left\{\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right\} = 0.5 \quad Gini(97 +) = 1 - \left\{\left(\frac{0}{4}\right)^2 + \left(\frac{4}{4}\right)^2\right\} = 0$
- $Gini(97) = \frac{6}{10}(0.5) + \frac{4}{10}(0) = 0.3$

- درآمد 97 دارای کمترین مقدار جینی و در نتیجه بیشترین بهره اطلاعاتی می‌باشد.


درآمد
< 97
>= 97

هرس کردن درخت تصمیم:

- ✓ در بعضی از درخت‌های تصمیم به دلیل وجود اغتشاش در داده‌ها از جمله حضور داده‌های پرت، تعدادی از شاخه‌ها به طور غیر متعارف رشد می کنند. از این رو روش‌هایی برای هرس کردن درخت تصمیم به وجود آمده است.
- ✓ این الگوریتم ها اغلب از محاسبات آماری برای حذف شاخه‌های غیرقابل اعتماد استفاده می‌کنند وبا حذف شاخه‌های غیرعادی درخت را کوچک تر و ساده تر می‌کنند و به موجب آن راحتتر می‌توان آن‌ها را فهمید در این صورت تصمیم‌گیری راحت‌تر خواهد بود.
- ✓ درخت‌های هرس شده سریع‌تر و با دقت بیشتر نسبت به درخت‌هایی که هرس نشده‌اند، می‌توانند داده‌های مستقل(چندتایی‌های جدیدی که در مجموعه آموزشی نبوده‌اند و هدف دسته‌بندی آن‌هاست) را دسته‌بندی کنند

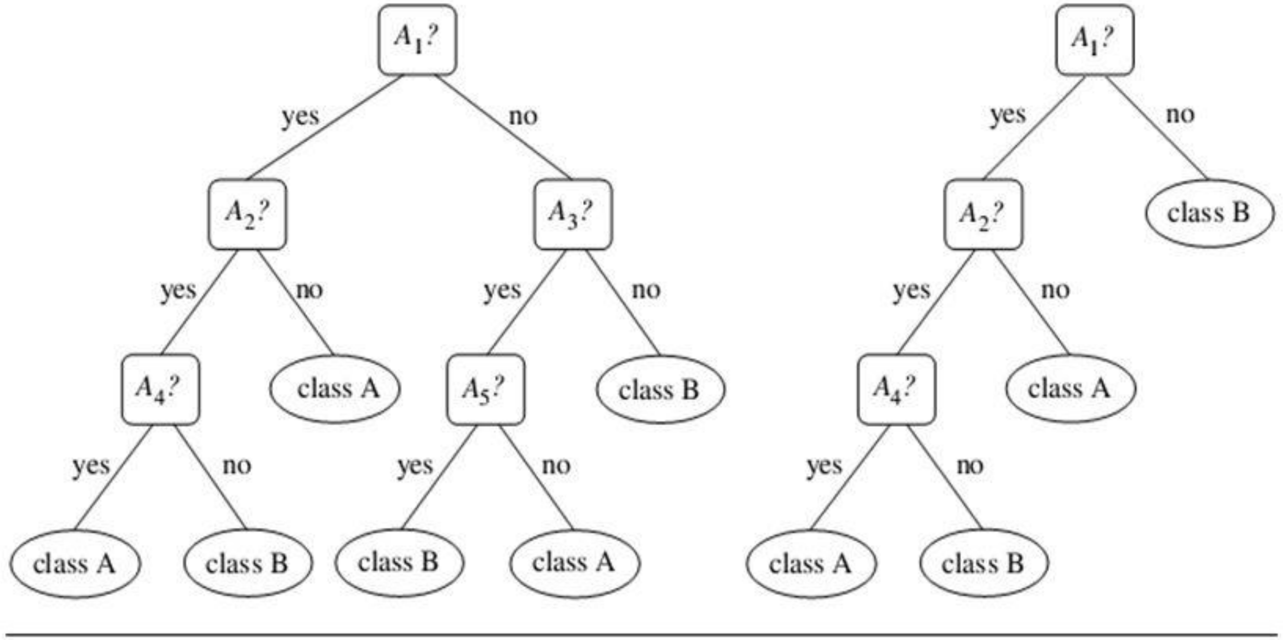
❖فرآیند هرس کردن درخت‌های تصمیم چگونه کار می‌کند؟

✓دو روش معمول، **هرس قبل از فرآیند شکل‌گیری درخت** و **هرس پس از شکل‌گیری درخت** است.

✓در روش هرس قیل از شکل‌گیری درخت، درخت با جلوگیری از گسترش بیشتر شاخه‌ها، هرس می‌شود.(چندتاییی‌هایی که در یک گروه قرار دارند بنابر شرایط، دیگر تقسیم نمی‌شوند و شاخه جدیدی ایجاد نمی‌شود).

✓با جلوگیری از گسترش شاخه از گره‌ای، آن گره به برگ تبدیل می‌شود و با توجه به دسته غالبی از چندتایی‌ها که درون برگ قرار گرفته‌اند، برچسب آن زده می‌شود.

Overfitting and Tree Pruning



An unpruned decision tree and a pruned version of it.

رگرسیون لجستیک (Logistic regression)

- یکی از تکنیک‌های پیشرفته آماری که در **« یادگیری ماشین»** در قسمت دسته بندی بسیار کاربرد دارد، **« رگرسیون لجستیک »** است.
- اغلب برای نمایش رابطه بین دو متغیر از مدل رگرسیونی کمک می‌گیریم.

در این صورت یک الگو برای پیش بینی متغیر وابسته **Y** براساس متغیر مستقل **X** ایجاد می شود. ولی باید توجه داشت که در مدل خطی ایجاد شده، هر دو متغیر مستقل و وابسته، کمی هستند.

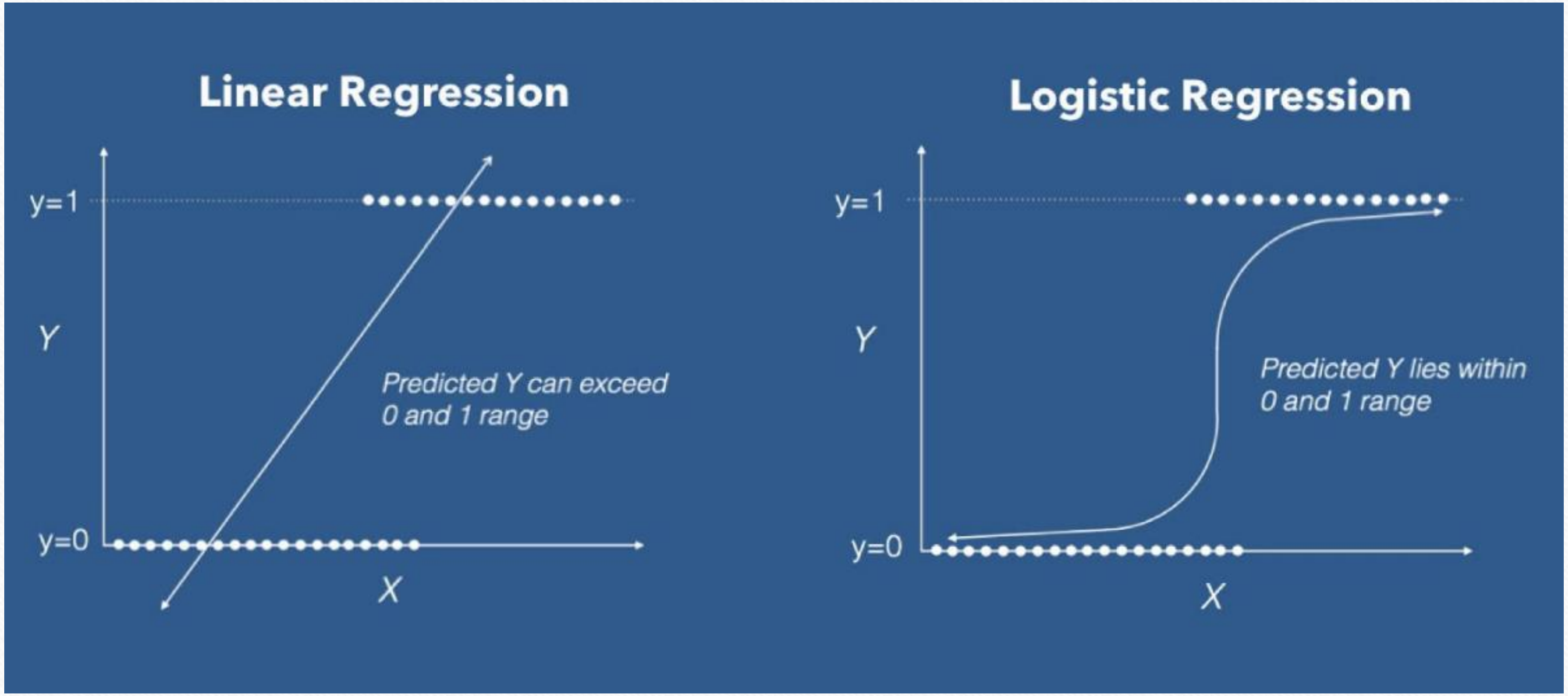
- همچنین شرط پیوسته بودن این مقدارها نیز در روش رگرسیون نهفته است. ولی ممکن است بخواهیم رابطه بین یک متغیر مستقل (با مقادیر پیوسته) را با یک متغیر وابسته با مقدارهای کیفی بسنجیم.

✓ در این حالت روش عادی رگرسیون خطی پاسخگو نخواهد بود و باید از « رگرسیون لجستیک » استفاده کرد

کاربرد رگرسیون لوجستیک

- از رگرسیون لجستیک بخصوص در زمینه‌های پزشکی، روانشناسی و علوم اجتماعی بسیار کمک گرفته می شود.
- برای مثال بررسی و ایجاد مدل رابطه بین میزان فعالیت روزانه و ابتلا به بیماری قند یک نمونه از تحلیل‌هایی است که در آن از مدل رگرسیون لجستیک کمک می‌گیرند.
- در این حالت متغیر مستقل، فعالیت روزانه با مقدارهای کمی است و متغیر وابسته کیفی نیز ابتلا یا عدم ابتلا به بیماری قند است که دارای دو مقدار ۰ و یا ۱ خواهد بود.
- همچنین در تحلیل حافظه انسان و رابطه آن با میزان خواب، روانشناسان آزمایشی را انجام می دهند که براساس مقدار ساعات متفاوت خواب افراد، یادآوری یا فراموشی کلمه ای را می‌سنجند. در این حالت میزان خواب متغیر مستقل با مقدارهای کمی پیوسته و متغیر وابسته کیفی با دو مقدار ۰ به معنی فراموشی و ۱ به معنی یادآوری صحیح است.

- ✓ این تکنیک در ابتدا در حوزه سلامت و برای بررسی بیماری یا عدم وجود بیماری به کار برده شد.
- ✓ در مطالعات بر روی داده‌های **کیفی دو وجهی**، متغیرپاسخ را می‌توان به صورت مقادیر 0 و 1 کد گذاری کرد. اگر برای چنین حالتی الگوی سنتی رگرسیونی به کار برده شود، یک طرف معادله فقط اعداد 0 و 1 را می‌تواند بپذیرد در صورتی که طرف دیگر تساوی از نظر تئوری بی‌نهایت مقدار را شامل می‌شود.



- راه حل برای این مسئله پیشنهاد می‌شود این است که طرف اول معادله را نیز به مقادیر **پیوسته** تبدیل کنیم. در این تکنیک به جای پیشگویی مقادیر متغیر پاسخ، احتمال این را که متغیر پاسخ یکی از مقادیر 0 یا 1 را بگیرد، محاسبه می‌شود.
- به عبارتی در این مدل پیشنهادی به جای استفاده از Y، **تابعی از احتمال Y**، به عنوان متغیر پاسخ به کار برده می‌شود. این کار در سه مرحله انجام می‌شود.

✓ **گام اول:** در این مرحله احتمال Y، که در معادله رگرسیون به جای Y است، محاسبه می‌شود تا به عنوان متغیرپاسخ وارد مدل شود در این صورت سمت چپ معادله رگرسیونی عددی بین 0 و 1 است.

$$P(Y=1)=\pi, \qquad P(Y=0)=1-\pi$$

✓ **گام دوم:** این مرحله شامل به کار بردن تابعی از احتمال Y در معادله است. این تابع **نسبت بخت (Odds Ratio)** نامیده می‌شود. در این مرحله سمت چپ معادله رگرسیونی عددی بین 0 و $+\infty$ خواهد شد.

$$OddsRatio = OR = \left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \frac{\pi}{1-\pi}$$

✓ **گام سوم:** این مرحله شامل استفاده از لگاریتم طبیعی تابع نسبت بخت به عنوان متغیرپاسخ در معادله است. در این صورت سمت چپ معادله مقداری بین $-\infty$ و $+\infty$ خواهد بود.

$$Ln(OR) = Ln\left(\frac{\pi}{1-\pi}\right)$$

- رگرسیون لجستیک برای ما این امکان را فراهم آورده است که متغیر وابسته دو وجهی را بر مبنای متغیرهای پیشگو یا مستقل پیش‌بینی کنیم. در این روش لگاریتم بخت متغیرهای وابسته، به عنوان ترکیب خطی از متغیرها در مدل قرار می‌گیرد.

$$\boldsymbol{Ln}\left(\frac{\boldsymbol{\pi}}{\boldsymbol{1}-\boldsymbol{\pi}}\right)=\boldsymbol{\beta}_{\boldsymbol{0}}+\boldsymbol{\beta}_{\boldsymbol{1}}\boldsymbol{X}_{\boldsymbol{1}}+\cdots+\boldsymbol{\beta}_{\boldsymbol{n}}\boldsymbol{X}_{\boldsymbol{n}}$$

- $\mathrm{Ln}\left(\frac{P(y_i=1)}{P(y_i=0)}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \rightarrow \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$
- با در نظر گرفتن

$P(y_i = 1) = \pi_1$ و $P(y_i = 0) = \pi_0$

$$\pi_1 = \frac{exp(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki})}{1+\exp(\beta_0+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki})}$$

نسبت بخت (Odds Ratio):

- برای آشنایی با مفهوم بخت به یک مثال می‌پردازیم.
- فرض کنید در یک خانواده که دارای شش فرزند هستند، 2 فرزند پسر و 4 دختر پس نسبت پسرها به دخترها برابر است با 2/4 ، این نسبت نشان می‌دهد که تعداد دخترها در این خانواده دو برابر تعداد پسرها است. از طرفی می دانیم در چنین خانواده ای احتمال انتخاب یک پسر از بین فرزندان برابر با $\frac{1}{3}$ و چنین احتمالی نیز برای دخترها برابر با $\frac{2}{3}$ است. حال اگر پیشامد **A** را انتخاب یکی از پسرها در بین فرزندان در نظر بگیریم، بخت یا شانس برایچنین پیشامدی برابر است با:

$$\frac{\frac{1}{3}}{1-\frac{1}{3}}=\frac{1}{2}$$

و این شانس برای دختر بودن برابر: $\frac{\frac{2}{3}}{1-\frac{2}{3}}=2$

پس بخت انتخاب دختر 2 برابر پسر هست.

- از طرفی بخت می تواند مقداری بیشتر از یک را اختیار کند.

- به منظور برآورد پارامترهای مدل، می‌توان از «تبدیل لوجیت» (Logit Transformation)استفاده کرد. این تبدیل را روی بخت $\frac{P(X)}{1-P(X)}$ که قبلا بیان شده، اجرا می‌کنیم. در این صورت رابطه را می‌توان به شکل زیر نوشت:

$$g(x) = ln\left(\frac{P(X)}{1-P(X)}\right) = \frac{\frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}}}{1-\frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}}} = \ln\left(e^{b_0+b_1x}\right) = b_0 + b_1x$$

- با استفاده از تابع درستنمایی و حداکثر سازی آن می‌توان مدل را براساس برآورد پارامترها به دست آورد. با این کار به یک دستگاه معادلات می‌رسیم که متاسفانه برای حل آن روش تحلیلی وجود ندارد و باید به کمک روش‌های عددی برآورد را انجام داد. خوشبختانه نرم‌افزارهای زیادی از جمله SPSS قادر هستند که محاسبات و برآوردهای مربوط به رگرسیون لوجستیک را انجام دهند و پارامترهای b_0 و b_1 را محاسبه کنند.

- در رگرسیون لوجستیک مقیاس اندازه‌گیری متغیرهای پیشگو هم می‌تواند **کمی**(سطوح اندازه‌گیری نسبی یا فاصله‌ای) و هم **کیفی** (سطوح اندازه‌گیری اسمی و ترتیبی) باشد.
- در الگو رگرسیون لوجستیک نیاز نیست متغیرهای مستقل توزیع نرمال داشته باشند.(برخلاف تحلیل تشخیصی) ولی در صورتی که توزیع نرمال چندمتغیره برای متغیرهای پیشگوی پیوسته که در مدل حضور ندارند برقرار باشند، برازش بهتری برای مدل میسر خواهد شد. تفسیر نتایج رگرسیون لوجستیک نیز آسان‌تر از نتایج تحلیل تشخیصی است.
- در صورتی که متغیرهای پیشگو مدل دارای **چندهمخطی** باشند، برآوردهای مدل اریب و خطای استاندارد مدل دارای نوسان خواهد بود.
- در مدل رگرسیون لوجستیک قواعد خاصی برای حجم نمونه و تعداد متغیرهای پیشگو تعریف نشده است. در برخی منابع حداقل حجم نمونه باری رگرسیون لوجستیک 50 و برخی منابع 100 تعیین شده است.

✓ **مثال)** فرض کنید تحقیقی در مورد تأثیر مصرف سیگارو الکل بر روی داشتن بیماری قلبی صورت گرفته است. اطلاعات در جدل زیر آمده است:

	1	2	3	4	5	6	7	8	9
واحد سیگار	7	6	1	7	5	2	0	6	0
واحد الکل	8	2	3	4	5	6	7	7	8
داشتن بیماری قلبی	بله	بله	بله	بله	بله	بله	خیر	خیر	خیر

$$\begin{cases} y_i = 0,1 \\ P(Y_i = 1) = p \end{cases} \rightarrow y_i \sim B(n,p)$$

$$f(y_i) = p^{y_i}(1-p)^{\,1-y_i}$$

$$\prod_{i=1}^{n} f(y_i) = p^{\sum_{i=1}^{n} y_i}(1-p)^{n-\sum_{i=1}^{n} y_i}$$

$$\ln\left(\prod_{i=1}^{n} f(y_i)\right) = \sum_{i=1}^{n} y_i \ln(p) + n - \sum_{i=1}^{n} y_i \ln(1-p)$$

- ما در اینجا فقط y را داریم که دارای توزیع برنولی با احتمال موفقیت p است.

$$p = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

به صورت زیر محاسبه می‌شود.$L(\alpha,\beta)$در این تابع درستنمایی

$$\begin{cases} \dfrac{\partial \ln L(\alpha,\beta)}{\partial \alpha} = 0 \\ \dfrac{\partial \ln L(\alpha,\beta)}{\partial \beta} = 0 \end{cases}$$

که $\hat{\alpha}$ و $\hat{\beta}$ با کمک روشهای تکرار عددی محاسبه می‌شوند.

نتایجی حاصل شده با کمک سیستم به صورت زیر است.

ضریب مدل	$\hat{\beta}$	SE	Wald	df	Sig	exp
مقدار ثابت α	-1.39				0.00	0.25
سیگار β_1	2.26				0.00	9.62
الکل β_2	-0.06				0.36	9.92

- $0.36 > 0.05 \rightarrow AH_0$

$$\begin{cases} H_0: \beta_1 = 0 \\ H_1: \beta_1 \neq 0 \end{cases} \rightarrow Sig = 0.00 < 0.05 \Rightarrow RH_0$$

$$\begin{cases} H_0: \beta_2 = 0 \\ H_1: \beta_2 \neq 0 \end{cases} \rightarrow Sig = 0.36 > 0.05 \Rightarrow AH_0$$

- ✓ پس متغیر الکل بایستی از مدل خارج شود اما مدل در حالت کلی به صورت زیر است:

$$logit(p) = -1.39 + 2.26x_1 - 0.06x_2$$

- ✓ که بعد از خارج کردن x_2 از مدل داریم:

$$logit(p) = -1.39 + 2.26x_1$$

- **سوال:** اگر فرد به تعداد یک واحد سیگار بیشتر مصرف کند، بخت ابتلا به بیماری قلبی افزایش می‌یابد؟

$$Odds = \frac{p}{1-p} = e^{-1.39+2.26(1)} = e^{2.26} = 9.62 \Rightarrow p = \frac{9.62}{1+9.62} \cong 0.9$$

✓ اگر سیگار ثابت باشد بخت الکل چقدر است؟ (پیش‌بینی در صورت ثابت بودن بقیه متغیرها)

✓ فقط x یک واحد افزایش کی‌یابد که به ستون exp مراجعه می‌کنیم.

انواع رگرسیون لجستیک:

- رگرسیون لجستیک باینری یا دو سطحی Binary logistic regression
- رگرسیون لجستیک چند سطحی Multinomial logistic regression
- رگرسیون لجستیک ترتیبی Ordinal logistic regression