دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر

تمر ین دوم درس یادگیری ماشین

دکتر ناظرفرد

غلامرضا دار ۴۰۰۱۳۱۰۱۸

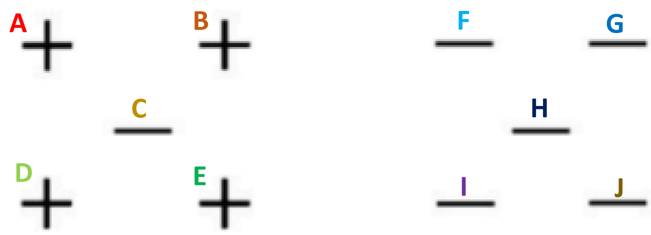
پاییز ۱۴۰۰

فهرست مطالب

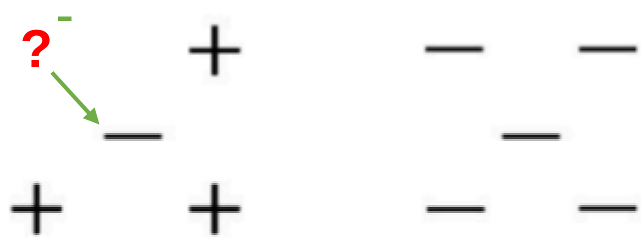
بخش اول: پرسشهای تشریحی

سوال ۱)

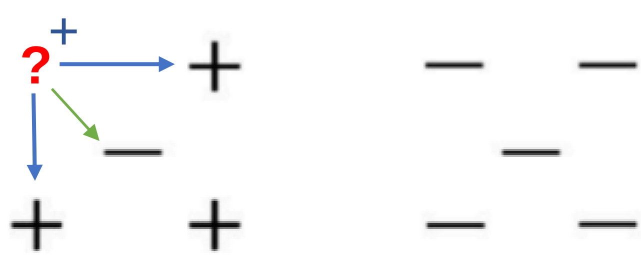
الف)



برای آسان تر شدن مسئله به هر کدام از نقاط دیتاست یک اسم میدهیم. در ادامه قرار است ۱۰ بار، هر بار یکی از این نقاط را به عنوان دیتای تست در نظر بگیریم (علامت سوال بگذاریم). این کار را برای مقادیر مختلف K انجام میدهیم. بهترین مقدار برای K مقداری است که کمترین میزان دسته بندیِ اشتباه را داشته باشد.



شکل ۱- با فرض K=۱ و حذف کردن داده A به عنوان داده تست



شکل ۲- با فرض K=3 و حذف کردن داده A به عنوان داده تست

در شکل های قبل دیدیم که با فرض داده A به عنوان بخش test در روش LOOCV و با فرض K=1 کلاس منفی به A نسبت داده شد که اشتباه است. اما با فرض K=3 چون $\frac{2}{3}$ همسایه های A مثبت بودند کلاس مثبت به A نسبت داده شد پس میتوان گفت برای A مقدار K=3 مناسب تر است.

نکته: چون تعداد کلاس های این مسئله ۲ است(مثبت و منفی) بهتر است مقادیر **فرد** را برای K در نظر بگیریم تا حالت برابر بودن تعداد همسایه های منفی و مثبت رخ ندهد.

Data / K	A	B	C	D	E	F	G	H	I	J	Total Error
1	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	5
3	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	1
5	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	1
7	✗	✗	✗	✗	✗	✓	✓	✓	✓	✓	5
9	✗	✗	✓	✗	✗	✓	✓	✓	✓	✓	4

همانطور که در جدول مشاهده میشود، مقادیر K=5 و K=3 کمترین میزان خطا را در حالت های مختلف داشتند. بنابر این بهترین مقدار K عدد ۵ یا ۳ است.

ب) راهی که در بخش الف استفاده شد. به ازای مقادیر مختلف K و با کمک Cross validation خطای مدل به ازای آن K را بدست می‌آوریم. بهترین K آن K ای است که کمترین خطا را به ازای نقاط مختلفِ داده داشته باشد.

سوال ۲)

هر دو Discriminative هستند زیرا به طور مستقیم $P(W|X)$ یا همان Posterior Probability را یاد میگیرند و اطلاعی از توزیع داده ها ندارند. در واقع $P(X|W)P(W)$ را نمیدانند و این باعث میشود نتوانند داده جدیدی تولید کنند.

در مقابل این دو مدل، مدل Naïve Bayes یک مدل Generative است زیرا توزیع داده ها را یاد میگیرد و با کمک آن Posterior Probability را حساب میکند. از آنجایی که این مدل توزیع داده ها را دارد میتواند داده های جدیدی تولید کند.

سوال ۳)

الف)

$$\sigma(a) = \frac{1}{1+e^{-a}} = (1+e^{-a})^{-1}$$

$$\frac{d\sigma(a)}{da} = (-1)(1+e^{-a})^{-2}(-1\times e^{-a}) = \frac{e^{-a}}{(1+e^{-a})^2}$$

$$\frac{1}{1+e^{-a}} \times \frac{e^{-a}}{1+e^{-a}} = \sigma(a)\times\frac{e^{-a}}{1+e^{-a}} = \sigma(a)\times\frac{e^{-a}+1-1}{1+e^{-a}}$$

$$\sigma(a)\times\left(\frac{e^{-a}+1}{1+e^{-a}} - \frac{1}{1+e^{-a}}\right) = \sigma(a)\times\left(1-\frac{1}{1+e^{-a}}\right) = \boldsymbol{\sigma(a)\times\big(1-\sigma(a)\big)}$$

ب)

$$P(c_1|x) = \sigma(w^T x)$$

$$\frac{\partial}{\partial w}\sigma(w^T x) = \sigma(w^T x)\big(1-\sigma(w^T x)\big)x^T$$

$$L(w) = -\sum_{i=1}^{n}\big(y^i \log P(c_1|x^i)\big) + \big(1-y^i\big)\log P(c_0|x^i)$$

$$L(w) = -\sum_{i=1}^{n}\big(y^i \log P\big(c_1\big|x^i\big)\big) + \big(1-y^i\big)\log\big(1-P\big(c_1\big|x^i\big)\big)$$

$$L(w) = -\sum_{i=1}^{n}\big(y^i \log\sigma(w^T x^i)\big) + \big(1-y^i\big)\log\big(1-\sigma(w^T x^i)\big)$$

ج)

$$\frac{\partial L(w)}{\partial w} = -\sum_{i=1}^{n} y^i \frac{\partial}{\partial \mathrm{w}} \log \sigma(w^T x^i) + \left(1 - y^i\right) \frac{\partial}{\partial \mathrm{w}} \log\left(1 - \sigma(w^T x^i)\right)$$

$$I\ )\ y^i \frac{1}{\sigma(w^T X)} \sigma\left(w^T x^i\right)\left(1 - \sigma\left(w^T x^i\right)\right) x^i$$

$$= y^i \left(1 - \sigma\left(w^T x^i\right)\right) x^i$$

$$II\ )\ (1 - y^i) \frac{1}{\left(1 - \sigma(w^T x^i)\right)} \times -\sigma\left(w^T x^i\right)\left(1 - \sigma\left(w^T x^i\right)\right) x^i$$

$$= y^i \sigma\left(w^T x^i\right) x^i - \sigma\left(w^T x^i\right) x^i$$

$$\frac{\partial L(w)}{\partial w} = -\sum_{i=1}^{n} I + II$$

$$\frac{\partial L(w)}{\partial w} = -\sum_{i=1}^{n} y^i \left(1 - \sigma\left(w^T x^i\right)\right) x^i + \ y^i \sigma\left(w^T x^i\right) x^i - \sigma\left(w^T x^i\right) x^i$$

$$\frac{\partial L(w)}{\partial w} = -\sum_{i=1}^{n} y^i x^i - y^i \sigma\left(w^T x^i\right) x^i + \ y^i \sigma\left(w^T x^i\right) x^i - \sigma\left(w^T x^i\right) x^i$$

$$\frac{\partial L(w)}{\partial w} = -\sum_{i=1}^{n} y^i x^i - \sigma\left(w^T x^i\right) x^i$$

$$\frac{\partial L(w)}{\partial w} = -\sum_{i=1}^{n} (y^i - \hat{y}^{\,i})\, x^i$$

$$\frac{\partial L(w)}{\partial w} = \sum_{i=1}^{n} (\hat{y}^{\,i} - y^i)\, x^i$$

بخاطر اینکه سوال **منفی** Log Likelihood را خواسته.

د) در حالتی که مجموعه داده های دو کلاس خطی جداپذیر باشند بیش برازشی که رخ میدهد احتمالا به شکل یک انحراف به سمت یکی از کلاس هاست به این دلیل که ممکن است تعداد داده های دو کلاس یک اندازه نباشند و مدل حین یادگیری به یک کلاس وزن بیشتری بدهد.

سوال ۴)

الف)

Age	Income	Student	Credit	Buy
Youth	High	No	Fair	No
Youth	High	No	Excellent	No
Middle	High	No	Fair	Yes
Senior	Medium	No	Fair	Yes
Senior	Low	Yes	Fair	Yes
Senior	Low	Yes	Excellent	No
Middle	Low	Yes	Excellent	Yes
Youth	Medium	No	Fair	No
Youth	Low	Yes	Fair	Yes
Senior	Medium	Yes	Fair	Yes
Youth	Medium	Yes	Excellent	Yes
Middle	Medium	No	Excellent	Yes
Middle	High	Yes	Fair	Yes
Senior	Medium	No	Excellent	No

- X1 = (age = youth, income = high, student = yes, credit = fair)
- X2 = (age = senior, income = low, student = no, credit = excellent)
- X3 = (age = middle-aged, income = medium, student = no, credit = fair)

برای ساده تر شدن حل این مسئله جدول زیر را تولید میکنیم.

Feature	Value	Count (Buy = Yes)	Count (Buy = No)	Sum
Age	Youth	2	3	5
	Middle	4	0	4
	Senior	3	2	5
Income	Low	3	1	4
	Medium	4	2	6
	High	2	2	4
Student	No	3	4	7
	Yes	6	1	7
Credit	Fair	6	2	8
	Excellent	3	3	6
		Total = 9	Total = 5	Total = 14

X1 = (age = youth, income = high, student = yes, credit = fair)

رابطه Bayes :

$$P(Buy \mid X) = \frac{P(X \mid Buy)\, P(Buy)}{P(X)}$$

حل مسئله برای "***خریدن فرد X1***" :

$$P(Buy = Yes \mid X = X1) = \frac{P(X = X1 \mid Buy = Yes)\, P(Buy = Yes)}{P(X = X1)}$$

با فرض مستقل بودن ویژگی ها :

$$= \frac{P(age = youth \mid\ Buy = Yes)\, P(income = high \mid\ Buy = Yes)\, P(student = Yes \mid\ Buy = Yes)\, P(credit = fair \mid\ Buy = Yes)\, P(Buy = Yes)}{P(age = youth)\, P(income = high)\, P(student = Yes)\, P(credit = fair)}$$

از روی جدول داریم :

- $P(age = youth) = \frac{5}{14} = 0.357$
- $P(income = high) = \frac{4}{14} = 0.285$
- $P(student = Yes) = \frac{7}{14} = 0.5$
- $P(credit = fair) = \frac{8}{14} = 0.571$

- $P(age = youth \mid\ Buy = Yes) = \frac{2}{9} = 0.222$
- $P(income = high \mid\ Buy = Yes) = \frac{2}{9} = 0.222$
- $P(student = Yes \mid\ Buy = Yes) = \frac{6}{9} = 0.666$
- $P(credit = fair \mid\ Buy = Yes) = \frac{6}{9} = 0.666$

$$P(Buy = Yes) = \frac{9}{14} = 0.642$$

با جاگذاری مقادیر در رابطه اصلی داریم :

$$\frac{0.222 * 0.222 *\ 0.666 * 0.666 *\ 0.642}{0.357 * 0.285 * 0.5 * 0.571} = \mathbf{0.483}$$

احتمال خریدن فرد X1 برابر 0.483 است.

X2 = (age = senior, income = low, student = no, credit = excellent)

رابطه Bayes :

$$P(Buy \mid X) = \frac{P(X \mid Buy)\, P(Buy)}{P(X)}$$

حل مسئله برای "***خریدن فرد X2***" :

$$P(Buy = Yes \mid X = X2) = \frac{P(X = X2 \mid Buy = Yes)\, P(Buy = Yes)}{P(X = X2)}$$

با فرض مستقل بودن ویژگی ها :

$$= \frac{P(age = senior \mid\ Buy = Yes)\, P(income = low \mid\ Buy = Yes)\, P(student = No \mid\ Buy = Yes)\, P(credit = excellent \mid\ Buy = Yes)\, P(Buy = Yes)}{P(age = senior)\, P(income = low)\, P(student = No)\, P(credit = excellent\ )}$$

از روی جدول داریم :

- $P(age = senior) = \frac{5}{14} = 0.357$
- $P(income = low) = \frac{4}{14} = 0.285$
- $P(student = No) = \frac{7}{14} = 0.5$
- $P(credit = Excellent) = \frac{6}{14} = 0.428$

- $P(age = senior \mid\ Buy = Yes) = \frac{3}{9} = 0.333$
- $P(income = low \mid\ Buy = Yes) = \frac{3}{9} = 0.333$
- $P(student = No \mid\ Buy = Yes) = \frac{3}{9} = 0.333$
- $P(credit = Excellent \mid\ Buy = Yes) = \frac{3}{9} = 0.333$

$$P(Buy = Yes) = \frac{9}{14} = 0.642$$

با جاگذاری مقادیر در رابطه اصلی داریم :

$$\frac{0.333 * 0.333 * 0.333 * 0.333 * 0.642}{0.357 * 0.285 * 0.5 * 0.428} = \mathbf{0.362}$$

احتمال خریدن فرد X2 برابر 0.362 است.

X3 = (age = middle-aged, income = medium, student = no, credit = fair)

رابطه Bayes :

$$P(Buy \mid X) = \frac{P(X \mid Buy)\, P(Buy)}{P(X)}$$

حل مسئله برای "***خریدن فرد X3***" :

$$P(Buy = Yes \mid X = X1) = \frac{P(X = X1 \mid Buy = Yes)\, P(Buy = Yes)}{P(X = X1)}$$

با فرض مستقل بودن ویژگی ها :

$$\frac{P(age = middleaged \mid\ Buy = Yes)\, P(income = medium \mid\ Buy = Yes)\, P(student = No \mid\ Buy = Yes)\, P(credit = fair \mid\ Buy = Yes)\, P(Buy = Yes)}{P(age = middleaged)\, P(income = medium)\, P(student = No)\, P(credit = fair)}$$

از روی جدول داریم :

- $P(age = middleaged) = \frac{4}{14} = 0.285$
- $P(income = medium) = \frac{6}{14} = 0.428$
- $P(student = No) = \frac{7}{14} = 0.5$
- $P(credit = fair) = \frac{8}{14} = 0.571$

- $P(age = middleaged \mid\ Buy = Yes) = \frac{4}{9} = 0.444$
- $P(income = medium \mid\ Buy = Yes) = \frac{4}{9} = 0.444$
- $P(student = No \mid\ Buy = Yes) = \frac{3}{9} = 0.333$
- $P(credit = fair \mid\ Buy = Yes) = \frac{6}{9} = 0.666$

$$P(Buy = Yes) = \frac{9}{14} = 0.642$$

با جاگذاری مقادیر در رابطه اصلی داریم :

$$\frac{0.444 * 0.444 * 0.333 * 0.666 *\ 0.642}{0.285 * 0.428 * 0.5 * 0.571} = \mathbf{0.805}$$

احتمال خریدن فرد X3 برابر 0.805 است.

ب) ابتدا Entropy(S) را محاسبه میکنیم. سپس Information gain همه Feature ها را محاسبه میکنیم. فیچری که بیشترین Information gain را داشته باشد را به عنوان راس درخت انتخاب میکنیم و با کمک آن دیتاست را افراز میکنیم.

$$E(S) = -\left(\frac{9}{14}\log\frac{9}{14} + \frac{5}{14}\log\frac{5}{14}\right) = 0.940$$

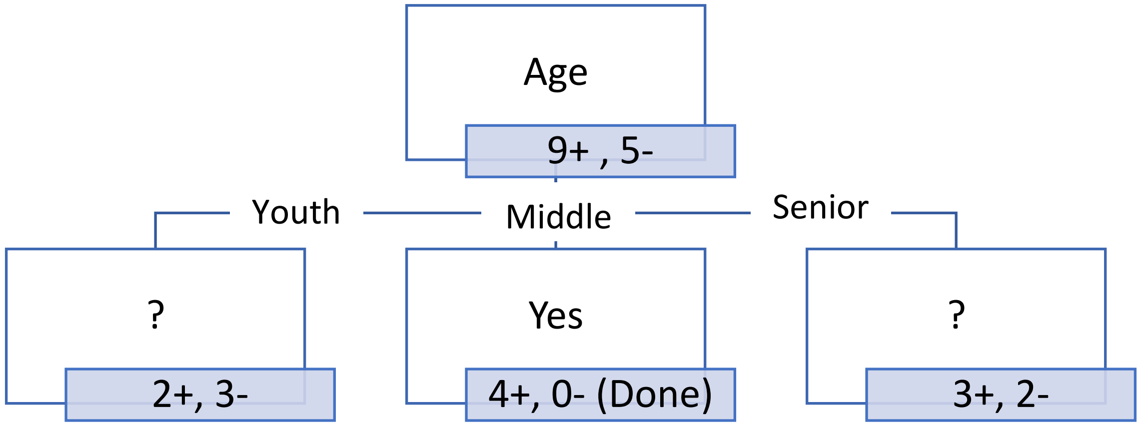
$$Gain(S, Age) \quad = 0.940 - \begin{pmatrix} \frac{5}{14}\left(-\left(\frac{2}{5}\log\frac{2}{5} + \frac{3}{5}\log\frac{3}{5}\right)\right) + \\ \frac{4}{14}\left(-\left(\frac{4}{4}\log\frac{4}{4} + 0\right)\right) + \\ \frac{5}{14}\left(-\left(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}\right)\right) \end{pmatrix} = \mathbf{0.246}$$

$$Gain(S, Income) = 0.940 - \begin{pmatrix} \frac{4}{14}\left(-\left(\frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4}\right)\right) + \\ \frac{6}{14}\left(-\left(\frac{4}{6}\log\frac{4}{6} + \frac{2}{6}\log\frac{2}{6}\right)\right) + \\ \frac{4}{14}\left(-\left(\frac{2}{4}\log\frac{2}{4} + \frac{2}{4}\log\frac{2}{4}\right)\right) \end{pmatrix} = 0.028$$

$$Gain(S, Student) = 0.940 - \begin{pmatrix} \frac{7}{14}\left(-\left(\frac{3}{7}\log\frac{3}{7} + \frac{4}{7}\log\frac{4}{7}\right)\right) + \\ \frac{7}{14}\left(-\left(\frac{6}{7}\log\frac{6}{7} + \frac{1}{7}\log\frac{1}{7}\right)\right) \end{pmatrix} = 0.151$$

$$Gain(S, Credit) \quad = 0.940 - \begin{pmatrix} \frac{8}{14}\left(-\left(\frac{6}{8}\log\frac{6}{8} + \frac{2}{8}\log\frac{2}{8}\right)\right) + \\ \frac{6}{14}\left(-\left(\frac{3}{6}\log\frac{3}{6} + \frac{3}{6}\log\frac{3}{6}\right)\right) \end{pmatrix} = 0.047$$

ویژگی Age بیشترین Information Gain را دارد پس در مرحله اول از این ویژگی برای تقسیم کردن دیتاست استفاده میکنیم.



Age	Income	Student	Credit	Buy
Youth	High	No	Fair	No
Youth	High	No	Excellent	No
Youth	Medium	No	Fair	No
Youth	Low	Yes	Fair	Yes
Youth	Medium	Yes	Excellent	Yes

در این بخش به دنبال بهترین ویژگی برای علامت سوال سمت چپ میگردیم.

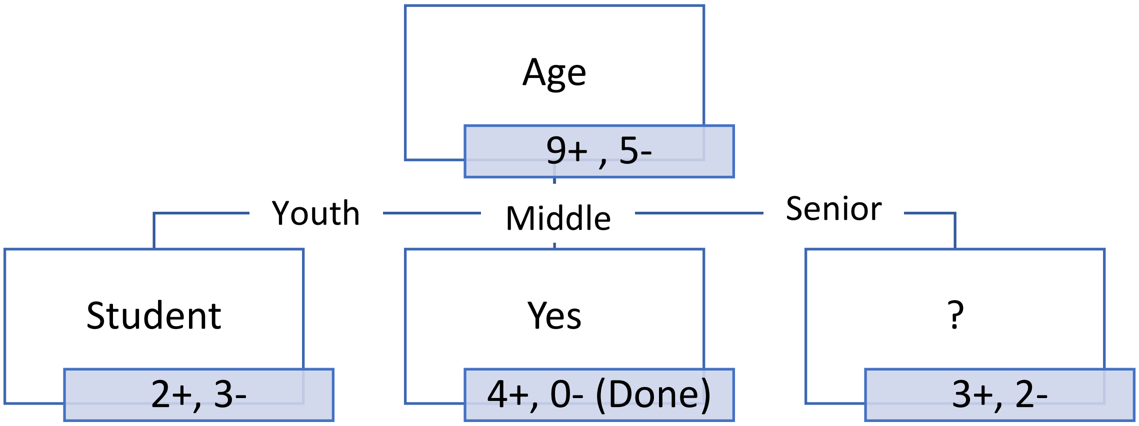
$$E(S) = -\left(\frac{3}{5}\log\frac{3}{5} + \frac{2}{5}\log\frac{2}{5}\right) = 0.970$$

$$Gain(S, Income) \ = 0.970 - \left(\begin{array}{c} \frac{1}{5}\left(-\left(0 + \frac{1}{1}\log\frac{1}{1}\right)\right) + \\ \frac{2}{5}\left(-\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right)\right) + \\ \frac{2}{5}\left(-\left(\frac{2}{2}\log\frac{2}{2} + 0\right)\right) \end{array}\right) = 0.570$$

$$Gain(S, Student) \ = 0.970 - \left(\begin{array}{c} \frac{3}{5}\left(-\left(\frac{3}{3}\log\frac{3}{3} + 0\right)\right) + \\ \frac{2}{5}\left(-\left(0 + \frac{2}{2}\log\frac{2}{2}\right)\right) \end{array}\right) = \mathbf{0.970}$$

$$Gain(S, Credit) \quad = 0.970 - \left(\begin{array}{c} \frac{3}{5}\left(-\left(\frac{2}{3}\log\frac{2}{3} + \frac{1}{3}\log\frac{1}{3}\right)\right) + \\ \frac{2}{5}\left(-\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right)\right) \end{array}\right) = 0.419$$

در این بخش ویژگی Student بیشترین Information Gain را دارد پس در این بخش از این ویژگی برای تقسیم کردن دیتاست استفاده میکنیم.



Age	Income	Student	Credit	Buy
Senior	Medium	No	Excellent	No
Senior	Low	Yes	Excellent	No
Senior	Low	Yes	Fair	Yes
Senior	Medium	No	Fair	Yes
Senior	Medium	Yes	Fair	Yes

در این بخش به دنبال بهترین ویژگی برای علامت سوال سمت راست میگردیم.

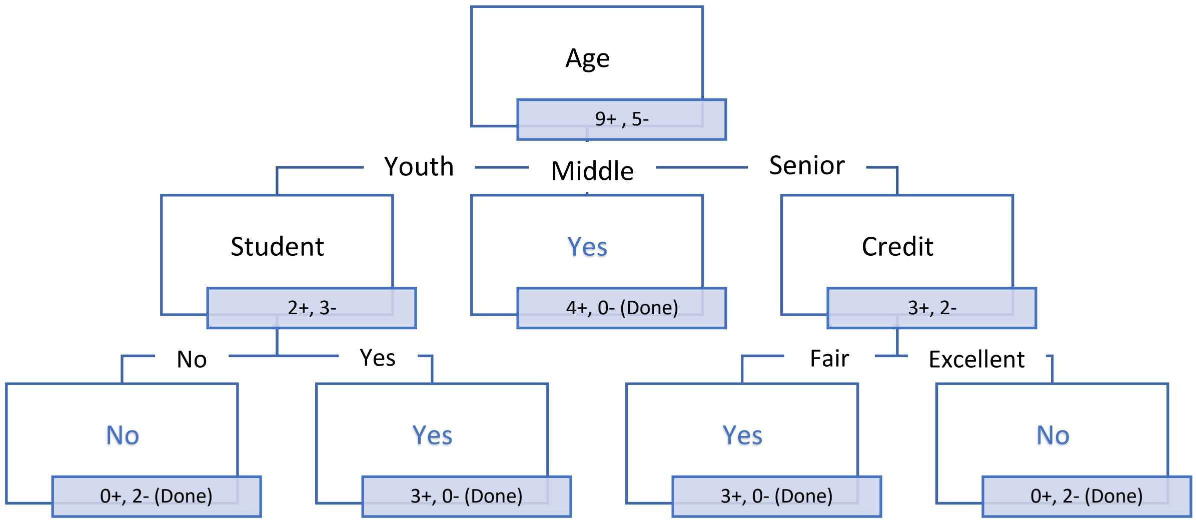
$$E(S) = -\left(\frac{2}{5}\log\frac{2}{5} + \frac{3}{5}\log\frac{3}{5}\right) = 0.970$$

$$Gain(S, Income) = 0.970 - \left(\begin{matrix}\frac{2}{5}\left(-\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right)\right) + \\ \frac{3}{5}\left(-\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right)\right)\end{matrix}\right) = 0.020$$

$$Gain(S, Student) = 0.970 - \left(\begin{matrix}\frac{2}{5}\left(-\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right)\right) + \\ \frac{3}{5}\left(-\left(\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right)\right)\end{matrix}\right) = 0.020$$

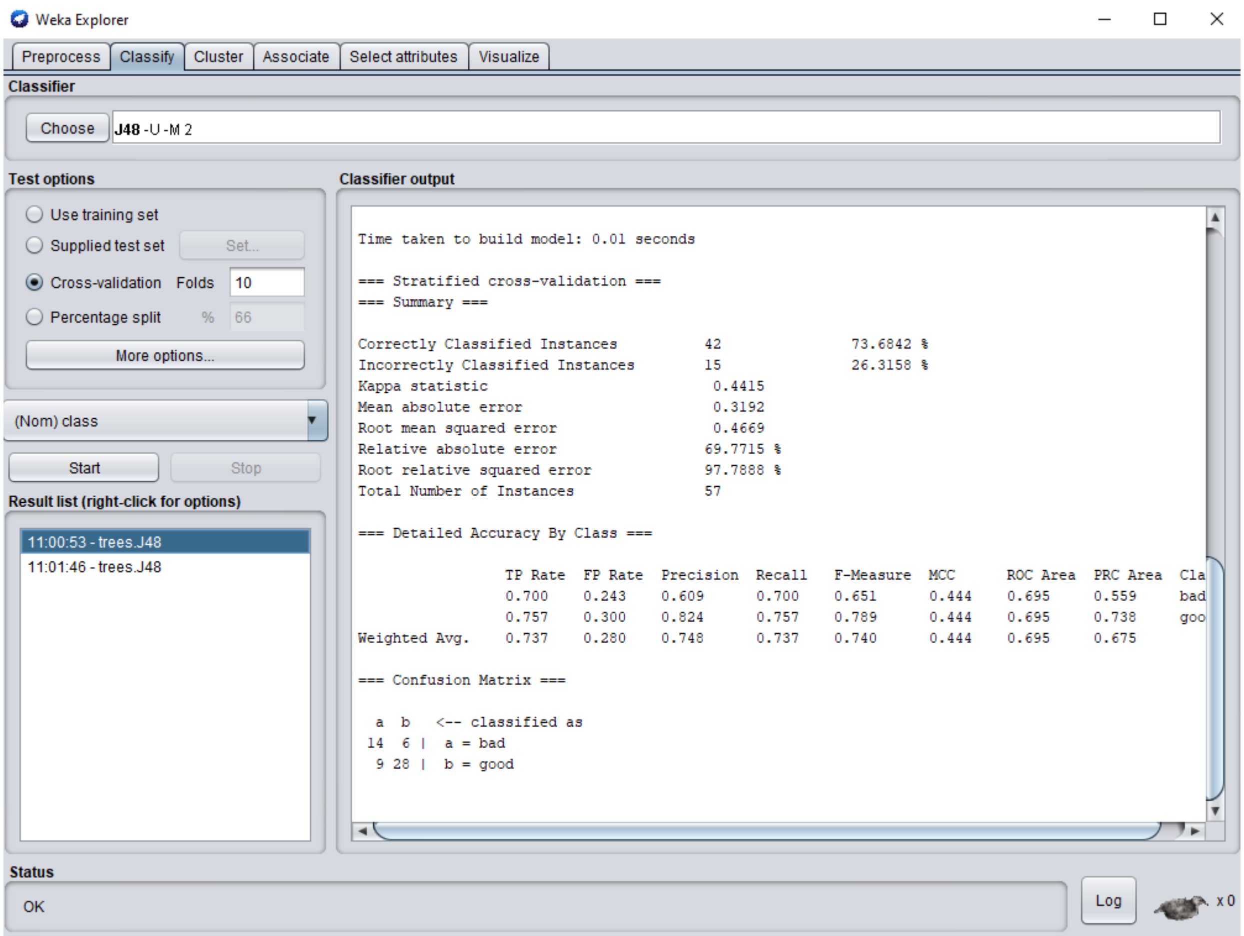
$$Gain(S, Credit) = 0.970 - \left(\begin{matrix}\frac{3}{5}\left(-\left(0 + \frac{3}{3}\log\frac{3}{3}\right)\right) + \\ \frac{2}{5}\left(-\left(\frac{2}{2}\log\frac{2}{2} + 0\right)\right)\end{matrix}\right) = \mathbf{0.970}$$

در این بخش ویژگی Credit بیشترین Information Gain را دارد پس در این بخش از این ویژگی برای تقسیم کردن دیتاست استفاده میکنیم.



سوال ۵)

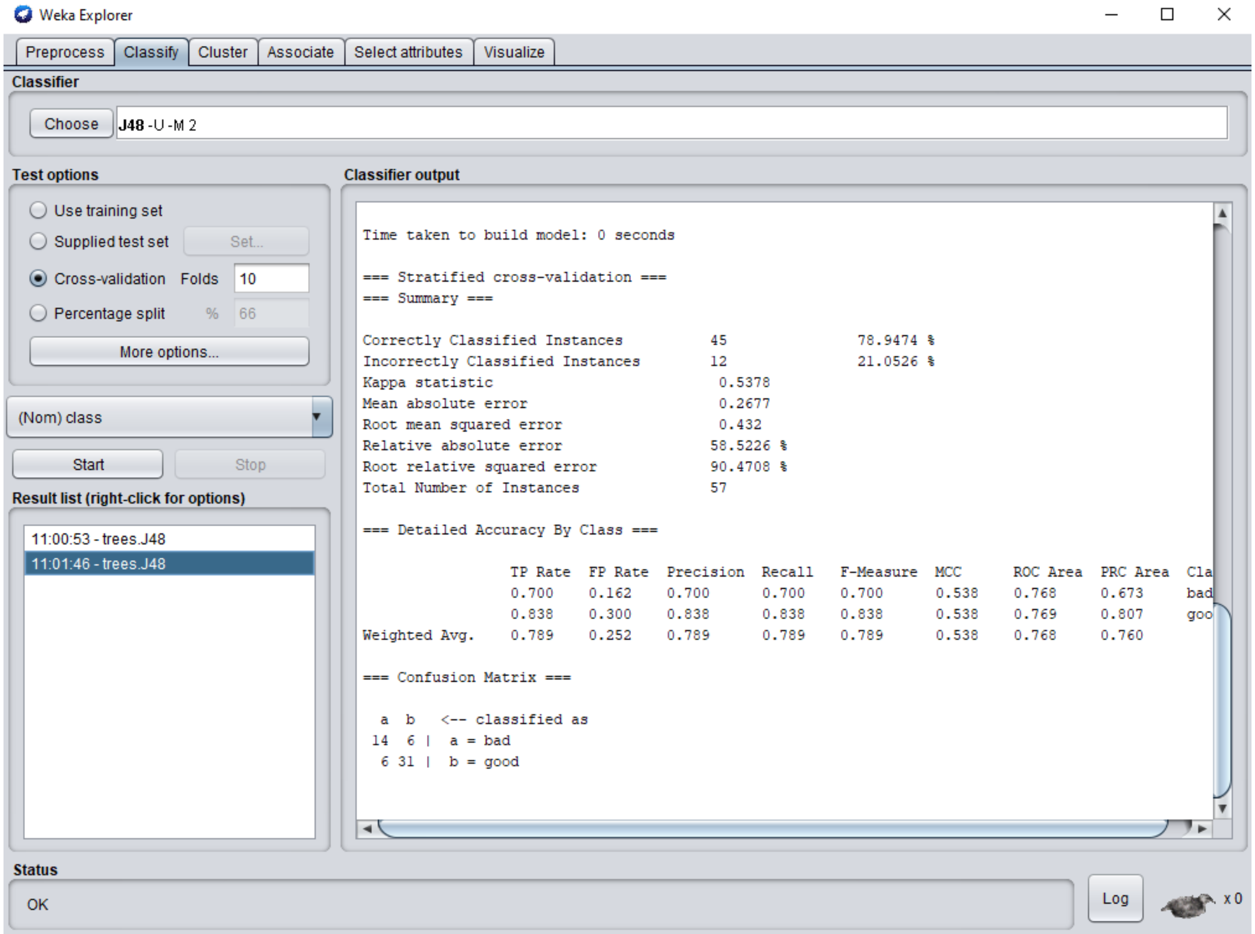
الف)



ب)

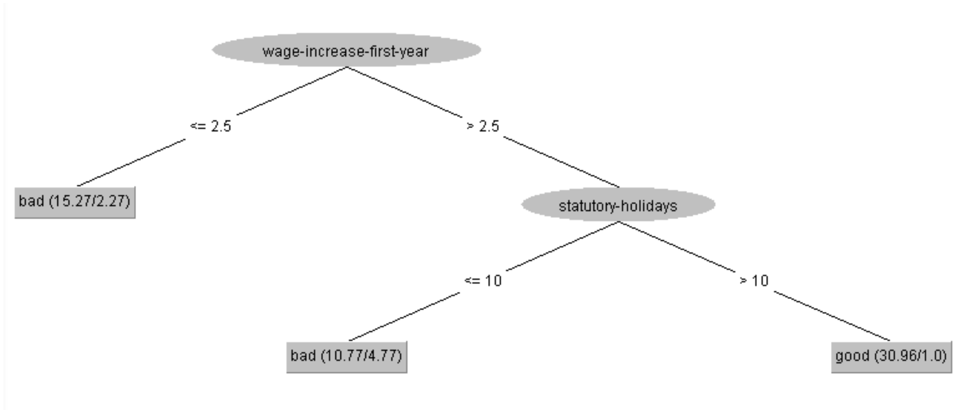
	Classified as Bad	Classified as Good
Bad	14	6
Good	9	28

ج)

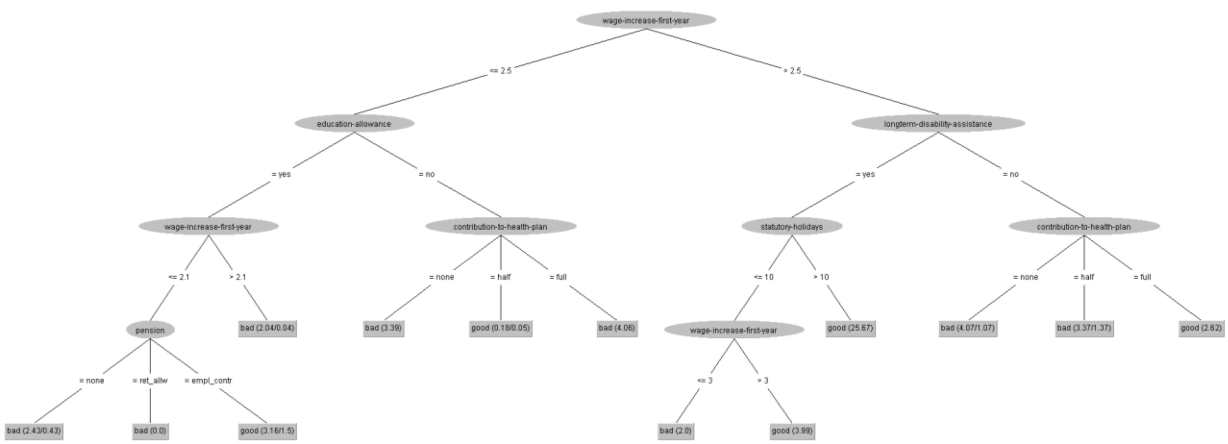


	Classified as Bad	Classified as Good
Bad	14	6
Good	6	31

د) تفاوتی که بین درخت تصمیم هرس شده و نشده دیده میشود در ارتفاع دو درخت و همچنین میزان تصمیم های اشتباه دو درخت است. درخت تصمیم هرس شده دارای اشتباهات بیشتری از درخت هرس نشده است اما احتمالا قابلیت **generalize** شدن بیشتری دارد.



شکل۳ - درخت تصمیم هرس شده



شکل۴ - درخت تصمیم هرس نشده

بخش دوم: پیاده سازی

سوال ۱)

نوت بوک مربوطه به این تمرین :

https://colab.research.google.com/drive/1hs3OGPk5St9bRO-bbWPr4TjBbgeGxlFn?usp=sharing

ابتدا داده ها را لود میکنیم. سپس لازم است دیتاست را shuffle کنیم و بعد از آن به نسبت 80-20 به بخش Train, Test تقسیم کنیم.

الف)

برای بخش الف سوال ویژگی های گسسته را مستقل از سایر ویژگی ها و ویژگی های پیوسته را تحت یک توزیع چند متغیره نرمال در نظر میگیریم. هدف از این سوال محاسبه Posterior Probability های زیر است:

$$p(heart\ disease \mid X)$$

$$p(!\, heart\ disease \mid X)$$

سپس با مقایسه کردن این دو احتمال به این نتیجه میرسیم که کدام outcome مناسب تر است و آن را به عنوان prediction برای دیتای X (فردX) در نظر میگیریم.

نکات قابل توجه در این سوال:

- بهتر است برای جلوگیری از float underflow به جای ضرب کردن احتمال های متعدد(که هر کدام ممکن است عدد بسیار کوچکی باشند) از posterior لگاریتم میگیریم تا این ضرب ها به جمع تبدیل شوند. لازم به ذکر است که این کار چون برای هر دو posterior انجام میشود و هدف مقایسه این دو است، تغییری در نتیجه ایجاد نمیکند.
- برای رفع مشکل احتمال0 برای برخی ویژگی ها، از Laplace smoothing استفاده میکنیم به این شکل که به هنگام محاسبه احتمال صورت و مخرج را با ۱ جمع میکنیم. این کار باعث میشود صفر شدن احتمال یک ویژگی باعث صفر شدن همه ترم های دیگر نشود.
- چون هر دو posterior بر عدد P(X) تقسیم میشوند، میتوانیم این تقسیم را نادیده بگیریم بدون آنکه در نتیجه تغییری ایجاد کنیم.

برای محاسبه $p(heart\ disease \mid X)$ به شکل زیر عمل میکنیم :

$$p(heart\ disease \mid X) = p(X \mid heart\ disease)\, p(heart\ disease)$$

و بخاطر مستقل بودن ویژگی های گسسته داریم :

$$p(X \mid heart\ disease) = \prod_{discrete\ i} p(xi \mid heart\ disease) \times \mathcal{N}(\mu, \Sigma)$$

که در آن $\mathcal{N}(\mu, \Sigma)$ توزیع چند متغیره گاوسی روی داده های با Label **مثبت** برای ویژگی های پیوسته است.

به طور مشابه برای $p(!\,heart\ disease \mid X)$ داریم:

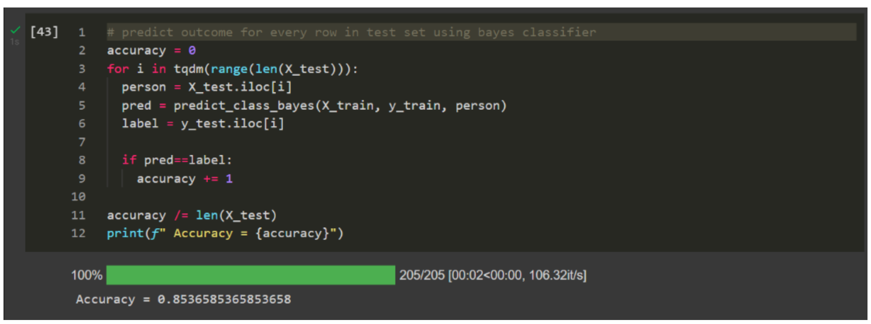
$$p(!\,heart\ disease \mid X) = \ p(X \mid !\,heart\ disease)\ p(!\,heart\ disease)$$

و بخاطر مستقل بودن ویژگی های گسسته داریم :

$$p(X \mid !\,heart\ disease) = \prod\nolimits_{discrete\ i} p(xi \mid !\,heart\ disease) \times \mathcal{N}(\mu, \Sigma)$$

که در آن $\mathcal{N}(\mu, \Sigma)$ توزیع چند متغیره گاوسی روی داده های با Label **منفی** برای ویژگی های پیوسته است.

پس از آموزش Bayes classifier روی داده های Train، کارایی مدل روی داده های Test ارزیابی شد که نتیجه، صحت ۸۵٪ بود.



```
# predict outcome for every row in test set using bayes classifier
accuracy = 0
for i in tqdm(range(len(X_test))):
  person = X_test.iloc[i]
  pred = predict_class_bayes(X_train, y_train, person)
  label = y_test.iloc[i]

  if pred==label:
    accuracy += 1

accuracy /= len(X_test)
print(f" Accuracy = {accuracy}")

100% 205/205 [00:02<00:00, 106.32it/s]
 Accuracy = 0.8536585365853658
```

شکل ۵ - آموزش و ارزیابی مدل *Bayes Classifier*

ب)

در این قسمت مشابه قسمت الف عمل میکنیم با این تفاوت که در این مثال همه ویژگی ها مستقل هستند نه فقط ویژگی های گسسته. بنابراین محاسبه $p(X \mid heart\ disease)$ به شکل زیر انجام می‌شود.

$$p(X \mid heart\ disease) = \prod\nolimits_{discrete\ i} p(xi\ |heart\ disease) \times \prod\nolimits_{continuous\ j} \mathcal{N}_j(\mu, \sigma)$$

که در آن $\mathcal{N}_j(\mu, \sigma)$ توزیع نرمال تک متغیری روی هر کدام از ویژگی های پیوسته با داده های label **مثبت** است.

به طور مشابه برای محاسبه $p(X \mid !\,heart\ disease)$ داریم:

$$p(X \mid !\,heart\ disease) = \prod\nolimits_{discrete\ i} p(xi \mid !\,heart\ disease) \times \prod\nolimits_{continuous\ j} \mathcal{N}_j(\mu, \sigma)$$

که در آن $\mathcal{N}_j(\mu, \sigma)$ توزیع نرمال تک متغیری روی هر کدام از ویژگی های پیوسته با داده های label **منفی** است.

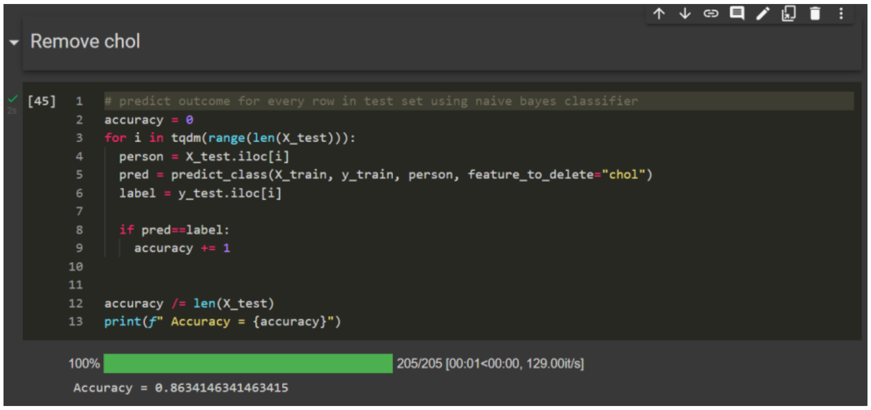
در این مورد به صحت ۸۶٪ رسیدیم.

```
# predict outcome for every row in test set using naive bayes classifier
accuracy = 0
for i in tqdm(range(len(X_test))):
  person = X_test.iloc[i]
  pred = predict_class(X_train, y_train, person)
  label = y_test.iloc[i]

  if pred==label:
    accuracy += 1


accuracy /= len(X_test)
print(f" Accuracy = {accuracy}")
```

```
100% 205/205 [00:02<00:00, 90.04it/s]
Accuracy = 0.8682926829268293
```

شکل ۶ - آموزش و ارزیابی مدل Naïve Bayes Classifier

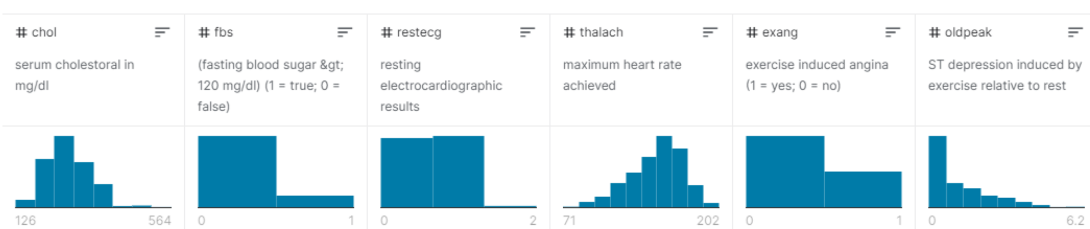
ج) نتیجه مدل قبل با حذف ویژگی های Chol, Oldpeak را به ترتیب در تصاویر زیر مشاهده میکنید.



شکل ۷- نتیجه نایو بیز پس از حذف ویژگی Chol

شکل ۸ - نتیجه نایو بیز پس از حذف ویژگی Oldpeak

همانطور که مشاهده میشود، حذف ویژگی Oldpeak بیشترین تاثیر را در صحت مدل می‌گذارد. دلیلی که ممکن است باعث این اتفاق شده باشد این است که مقادیر ویژگی Oldpeak اصلا از توزیع نرمال پیروی نمیکنند. با مراجعه به صفحه دیتاست در وبسایت Kaggle نیز میتوان این مورد را مشاهده کرد.



سوال ۲)

نوت بوک مربوطه به این تمرین :

https://colab.research.google.com/drive/1PQ_OsTzple1dd5EEmL0-kiUc9Ne5xfIR?usp=sharing

الف) ابتدا داده ها را load میکنیم و تعداد مقادیر گم شده هر ستون را پیدا میکنیم.

```
a = train_df.isna().sum()
a.loc[a>0]
```

```
Age          177
Cabin        687
Embarked       2
dtype: int64
```

```
a = test_df.isna().sum()
a.loc[a>0]
```

```
Age       86
Fare       1
Cabin    327
dtype: int64
```

So we have to deal with missing values from columns : "Age", "Cabin", "Embarked", "Fare"

همانطور که دیده میشود، ستون های Age, Cabin, Embarked, Fare نیاز به بررسی دارند. بقیه ستون ها مقدار گم شده ندارند. لازم است برای این ویژگی ها و یک سری ویژگی های دیگر تصمیماتی بگیریم. بعد از بررسی های انجام شده روی دیتاست تصمیم های زیر گرفته شد:

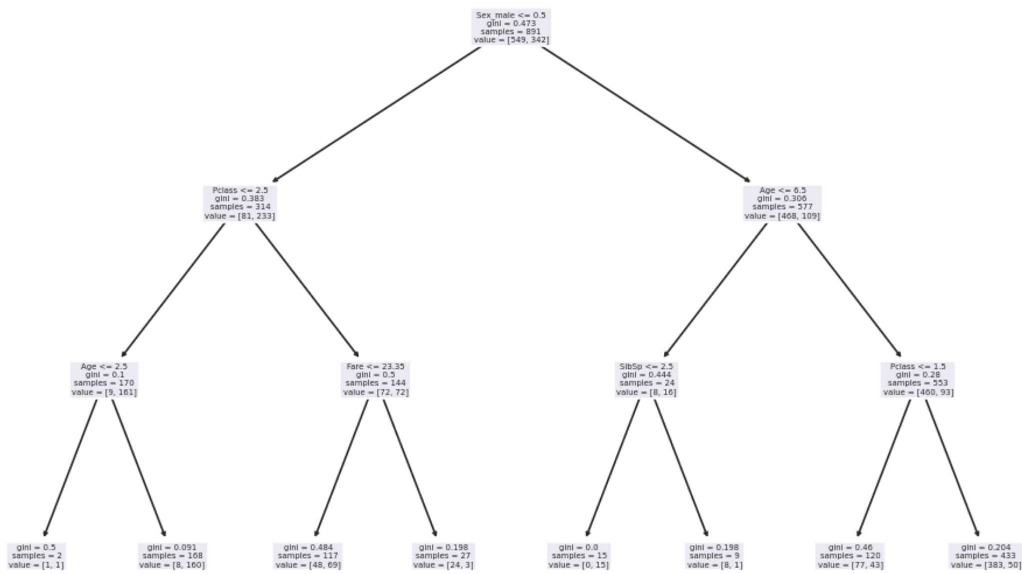
- **ویژگی Embarked :** جایگزین کردن مُد داده ها به جای مقادیر گم شده.
- **ویژگی Age :** جایگزین کردن مُد داده ها به جای مقادیر گم شده (مناسب تر از mean است چون outlier ها را در نظر نمیگیرد)
- **ویژگی Fare :** جایگزین کردن میانگین مقادیر با مقادیر گم شده.
- **ویژگی Name :** از آنجایی که اسم یک فرد تاثیری در زنده ماندن یا نماندن آن فرد ندارد ستون Name را حذف میکنیم.
- **ویژگی Cabin :** از آنجایی که تعداد بسیار زیادی از مقادیر این ویژگی گم شده است، جایگزین کردن میانگین میتواند مدل را گم راه کند و اطلاعات نادرستی به دیتاست اضافه کند. بنابراین این ویژگی را نیز حذف میکنیم.
- **ویژگی Ticket :** این ویژگی احتمالا میتواند پس از یک مرحله استخراج اطلاعات مفید واقع شود(مثلا تحلیل کنیم که اعداد و حروف های روی Ticket به چه معنا هستند) اما برای این سوال این ویژگی را حذف میکنیم.

پس از این مرحله نیاز است مقادیر Categorical را Tokenize کنیم تا به مقادیر عددی تبدیل شوند. ویژگی های Sex, Embarked از این دسته ویژگی ها هستند که به کمک روش one-hot encoding این داده ها را Tokenize میکنیم.

	PassengerId	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male	Embarked_C	Embarked_Q	Embarked_S
count	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000	891.000000
mean	446.000000	2.308642	28.566970	0.523008	0.381594	32.204208	0.352413	0.647587	0.188552	0.086420	0.725028
std	257.353842	0.836071	13.199572	1.102743	0.806057	49.693429	0.477990	0.477990	0.391372	0.281141	0.446751
min	1.000000	1.000000	0.420000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	223.500000	2.000000	22.000000	0.000000	0.000000	7.910400	0.000000	0.000000	0.000000	0.000000	0.000000
50%	446.000000	3.000000	24.000000	0.000000	0.000000	14.454200	0.000000	1.000000	0.000000	0.000000	1.000000
75%	668.500000	3.000000	35.000000	1.000000	0.000000	31.000000	1.000000	1.000000	0.000000	0.000000	1.000000
max	891.000000	3.000000	80.000000	8.000000	6.000000	512.329200	1.000000	1.000000	1.000000	1.000000	1.000000

- نمای کلی دیتاست پس از انجام عملیات های پیش پردازش۹شکل

ب) در این مرحله دیتاست ما آماده استفاده به عنوان دیتای ترین یک مدل DecisionTree است.



- درخت تصمیم تولید شده برای این مسئله۱۰شکل

امتیاز(score) اعلام شده توسط sklearn برای این درخت تصمیم برابر `0.827160493` بود.

ج) نتایج دسته بندی مسافران در دیتاست Test را میتوانید در این بخش مشاهده کنید.

در شکل ۷ خلاصه ای از افراد بازمانده را مشاهده میکنید.

	PassengerId	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male	Embarked_C	Embarked_Q	Embarked_S
1	893	3	47.0	1	0	7.0000	1	0	0	0	1
4	896	3	22.0	1	1	12.2875	1	0	0	0	1
6	898	3	30.0	0	0	7.6292	1	0	0	1	0
8	900	3	18.0	0	0	7.2292	1	0	1	0	0
12	904	1	23.0	1	0	82.2667	1	0	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...
409	1301	3	3.0	1	1	13.7750	1	0	0	0	1
410	1302	3	21.0	0	0	7.7500	1	0	0	1	0
411	1303	1	37.0	1	0	90.0000	1	0	0	1	0
412	1304	3	28.0	0	0	7.7750	1	0	0	0	1
414	1306	1	39.0	0	0	108.9000	1	0	1	0	0

150 rows × 11 columns

شکل ۱۱ - ۱۵۰ نفر از مسافران نجات یافتند

همچنین در این بخش میتوانیم ببینیم افراد بازمانده چه ویژگی هایی داشته اند یا مقادیر ویژگی های مختلف آنها در چه بازه ای یا دارای چه میانگینی بوده است که میتواند بسیار مفید باشد.

	PassengerId	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male	Embarked_C	Embarked_Q	Embarked_S
count	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000
mean	1095.406667	2.126667	28.450533	0.480000	0.506667	49.519917	0.960000	0.040000	0.273333	0.160000	0.566667
std	122.631608	0.892148	15.042604	0.672768	0.800559	74.076355	0.196616	0.196616	0.447164	0.367834	0.497196
min	893.000000	1.000000	0.170000	0.000000	0.000000	6.950000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	985.000000	1.000000	21.000000	0.000000	0.000000	8.553150	1.000000	0.000000	0.000000	0.000000	0.000000
50%	1092.500000	2.000000	24.000000	0.000000	0.000000	20.787500	1.000000	0.000000	0.000000	0.000000	1.000000
75%	1200.500000	3.000000	36.000000	1.000000	1.000000	54.856275	1.000000	0.000000	1.000000	0.000000	1.000000
max	1306.000000	3.000000	76.000000	4.000000	4.000000	512.329200	1.000000	1.000000	1.000000	1.000000	1.000000

شکل ۸ - خلاصه ای از ویژگی های افراد بازمانده

در شکل ۹ خلاصه ای از افراد نجات نیافته را مشاهده میکنید

	PassengerId	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male	Embarked_C	Embarked_Q	Embarked_S
0	892	3	34.5	0	0	7.8292	0	1	0	1	0
2	894	2	62.0	0	0	9.6875	0	1	0	1	0
3	895	3	27.0	0	0	8.6625	0	1	0	0	1
5	897	3	14.0	0	0	9.2250	0	1	0	0	1
7	899	2	26.0	1	1	29.0000	0	1	0	0	1
...	...	...	...	...	...	...	...	...	...	...	...
407	1299	1	50.0	1	1	211.5000	0	1	1	0	0
413	1305	3	21.0	0	0	8.0500	0	1	0	0	1
415	1307	3	38.5	0	0	7.2500	0	1	0	0	1
416	1308	3	21.0	0	0	8.0500	0	1	0	0	1
417	1309	3	21.0	1	1	22.3583	0	1	1	0	0

268 rows × 11 columns

شکل ۱۲ - ۲۶۸ نفر از مسافران نجات نیافتند

همچنین در این بخش میتوانیم ببینیم افراد نجات نیافته چه ویژگی هایی داشته اند یا مقادیر ویژگی های مختلف آنها در چه بازه ای یا دارای چه میانگینی بوده است که میتواند بسیار مفید باشد.

	PassengerId	Pclass	Age	SibSp	Parch	Fare	Sex_female	Sex_male	Embarked_C	Embarked_Q	Embarked_S
count	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000	150.000000
mean	1095.406667	2.126667	28.450533	0.480000	0.506667	49.519917	0.960000	0.040000	0.273333	0.160000	0.566667
std	122.631608	0.892148	15.042604	0.672768	0.800559	74.076355	0.196616	0.196616	0.447164	0.367834	0.497196
min	893.000000	1.000000	0.170000	0.000000	0.000000	6.950000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	985.000000	1.000000	21.000000	0.000000	0.000000	8.553150	1.000000	0.000000	0.000000	0.000000	0.000000
50%	1092.500000	2.000000	24.000000	0.000000	0.000000	20.787500	1.000000	0.000000	0.000000	0.000000	1.000000
75%	1200.500000	3.000000	36.000000	1.000000	1.000000	54.856275	1.000000	0.000000	1.000000	0.000000	1.000000
max	1306.000000	3.000000	76.000000	4.000000	4.000000	512.329200	1.000000	1.000000	1.000000	1.000000	1.000000

شکل ۸ - خلاصه ای از ویژگی های افراد نجات نیافته

سوال ۳)

نوت بوک مربوطه به این تمرین :

https://colab.research.google.com/drive/1oOhQ-PRDrJndli3CoxiARigyVg-83Wgv?usp=sharing

الف) در این بخش با استفاده از Cross validation با ۱۰ fold بهترین مقدار K را برای KNN پیدا میکنیم. روند کار به این صورت است که به ازای هر K ، ۱۰ بار مدل را ارزیابی میکنیم به شکلی که هر بار یک دهم دیتاست را به عنوان داده تست و ۹ دهم دیتاست را به عنوان داده ترین استفاده میکنیم. روند آموزش و ارزیابی مدل را در تصویر زیر مشاهده میکنید.

```
k = 1
test_wrongs : 0, 0, 1, 0, 0, 1, 0, 2, 0, 3,
train loss = 0.0407
test loss = 0.0467

k = 2
test_wrongs : 0, 0, 1, 0, 3, 1, 0, 2, 0, 1,
train loss = 0.0533
test loss = 0.0533

k = 3
test_wrongs : 0, 0, 1, 0, 1, 1, 0, 2, 0, 1,
train loss = 0.0385
test loss = 0.04

k = 4
test_wrongs : 0, 1, 1, 0, 3, 1, 0, 1, 0, 1,
train loss = 0.04
test loss = 0.0533

k = 5
test_wrongs : 0, 1, 0, 0, 2, 1, 0, 2, 0, 2,
train loss = 0.0356
test loss = 0.0533

k = 6
test_wrongs : 0, 1, 0, 0, 3, 1, 0, 1, 0, 0,
train loss = 0.0393
test loss = 0.04

k = 7
test_wrongs : 0, 1, 0, 0, 2, 1, 0, 1, 0, 1,
train loss = 0.0333
test loss = 0.04
```

شکل ۱۳ - روند آموزش مدل KNN

پس از اتمام آموزش نمودار K - avg_loss را رسم میکنیم. در این نمودار محور افقی مقدارK ، منحنی آبی Training_loss و منحنی نارنجی Validation_loss میباشد. این نمودار به ما کمک میکند بهترین مقدار K را پیدا کنیم. بهترین مقدار K مقداری است که کمترین validation loss را داشته باشد.

شکل ۱۴ - نمودار K - avg_loss برای پیدا کردن بهترین مقدار K

بهترین مقادیر K را به ترتیب در تصویر زیر مشاهده میکنید.

```
Top performing k:
K         Loss
13        0.03333333333333333
15        0.03333333333333333
6         0.039999999999999994
7         0.039999999999999994
8         0.039999999999999994
```

شکل ۱۵- جدول بهترین مقادیر K

ب) در نهایت برای مقدار K=13 ماتریس درهم ریختگی (Confusion Matrix) را رسم میکنیم.

```
Train Confusion Matrix
    0  1  2
0 [45, 0, 0]
1 [0, 45, 2]
2 [0, 1, 42]

Test Confusion Matrix
   0  1  2
0 [5, 0, 0]
1 [0, 3, 0]
2 [0, 1, 6]
```

منابع

تمرین پیاده سازی اول

https://colab.research.google.com/drive/1hs3OGPk5St9bRO-bbWPr4TjBbgeGxlFn?usp=sharing

تمرین پیاده سازی دوم

https://colab.research.google.com/drive/1PQ_OsTzple1dd5EEmL0-kiUc9Ne5xfIR?usp=sharing

تمرین پیاده سازی سوم

https://colab.research.google.com/drive/1oOhQ-PRDrJndli3CoxiARigyVg-83Wgv?usp=sharing

لینک گیتهاب

https://github.com/Gholamrezadar/machine-learning-exercises

Naïve Bayes

https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/#5naivebayesexamplebyhand