بخش تئوری

سوال ۱)

الف) پیوسته، کمی-نسبت

ب) پیوسته، کمی-بازه

ج) پیوسته، کمی-نسبت

د) گسسته، کیفی-ترتیبی

ح) پیوسته، کمی-بازه

ت) گسسته، کیفی-ترتیبی

سوال ۲)

الف) نویز داده را خراب می‌کند پس مطلوب نیست ولی تشخیص outlier می‌تواند همان چیزی باشد که ما در بین داده‌ها دنبال آن هستیم پس می‌تواند مطلوب باشد.

ب) بله، با توجه به اینکه نویز می‌تواند باعث ایجاد داده‌هایی نامعمول شود پس ممکن است این داده‌ها به عنوان outlier در بین داده‌های ما به نظر برسند.

ج) نه همیشه outlier نیستند چون ممکن است این اشیای نویز شبیه همان داده‌های ما باشند و لزومی ندارد که outlier باشند.

د) نه این‌طور نیست زیرا outlierها ممکن است که همین داده‌های ما باشند و به خاطر نویز به وجود نیامده باشند.

سوال ۳)

$$Euclidean \rightarrow \|X - Y\|_2$$

$$Correlation \rightarrow \frac{cov(x,y)}{\sigma_x \sigma_y}$$

$$Cosine \rightarrow \frac{\langle d_1, d_2 \rangle}{\|d_1\| \|d_2\|}$$

$$Jaccard \rightarrow \frac{f_{11}}{f_{11} + f_{10} + f_{01}}$$

الف)

$$Euclidean \rightarrow \sqrt{1+1+1+1} = 2$$

$$Correlation \rightarrow 0$$

$$Cosine \rightarrow 0$$

ب)

$$Euclidean \rightarrow \sqrt{9+4+1+4+0+4} = \sqrt{22}$$

$$Correlation \rightarrow 0$$

$$Cosine \rightarrow 0$$

ج)

$$Correlation \rightarrow \frac{1}{4}$$

$$Cosine \rightarrow \frac{1+1+1}{\sqrt{4} \times \sqrt{4}} = \frac{3}{4}$$

$$Jaccard \rightarrow \frac{3}{5}$$

سوال ۴)

الف) وقتی برخی ویژگی‌ها در بعضی تاپل‌ها مقدار ندارند، چند کار می‌توانیم بکنیم؛ مثلا، آن تاپل‌ها را حذف کنیم یا مثلا میانگین آن ویژگی در بقیه تاپل‌ها را برای آن‌ها جایگزین کنیم یا به عنوان راه دیگر می‌توانیم با توجه به مسئله‌ای که با آن رو به رو هستیم، مقداری را برای آن‌ها جایگزین کنیم که در نتیجه‌ی آن مسئله بی‌تاثیر باشند.

ب) وقتی با مجموعه داده‌ای رو به رو هستیم که ابعاد بسیار بالایی دارد باید از عملیات‌های کاهش ابعاد یا همان Dimensionality Reduction و به عنوان دو راه حل می‌توانیم از تکنیک‌های PCA (آنالیز مولفه اصلی) یا LDA (آنالیز تشخیص خطی) استفاده کنیم تا بتوانیم ویژگی‌های اصلی را شناسایی کنیم و فقط آن‌هایی که ارزش بیشتری دارند را مورد تحلیل و بررسی قرار دهیم.

ج) برای حل مشکل عدم توازن در داده‌ها مثلا می‌توانیم با دادن وزن به دسته‌های مختلف داده، این عدم توازن را کنترل کنیم یا به عنوان روش‌های دیگر می‌توانیم از Oversampling به معنی افزایش دسته‌ی کوچکتر با کمک کپی کردن داده‌های آن یا Undersampling به معنی حذف یک سری از داده‌ها از دسته‌ی بزرگتر برای کنترل این عدم توازن، استفاده کنیم.

د) برای حل مشکل بیش‌برازش می‌توانیم از روش‌هایی مثل توقف زود هنگام الگوریتم برای اینکه به بیش‌برازش نرسیم، استفاده کنیم؛ همچنین می‌توانی از Regularization استفاده کنیم که به نوعی یک جریمه به تابع هزینه اضافه می‌کند تا افزایش بیش از حد آن را کنترل کند و همچنین می‌توانیم با حذف و هرس گره‌هایی از درخت تصمیم که کمترین تاثیر را در نتیجه خروجی ما دارند نیز به ساده کردن مدل بپردازیم.

ه) در اینجا برخلاف قسمت قبل، مدل ما در مرحله آموزش بیش از حد ساده شده و نمی‌تواند نتیجه مناسبی به ما بدهد و خطای زیادی دارد که برای حل این مشکل باید مدل را پیچیده‌تر کنیم که می‌توانیم از روش‌هایی مثل افزایش پارامترهای مسئله یا افزایش لایه‌های یادگیری استفاده کنیم.

سوال ۵)

الف) بله رگرسیون خطی نسبت به outlierها حساس است و با افزایش تعداد outlierها، آن‌ها روی معادله‌ی رگرسیون خطی‌ای که میخواهیم پیدا کنیم تاثیر می‌گذارند و می‌توانند آن را تغییر دهند.

ب) برای اندازه‌گیری خطا در این الگوریتم معمولا از MSE استفاده می‌کنیم (میانگین مجذور فاصله‌ها). در این روش پس از تفریق مقدار نقطه‌ی پیش‌بینی شده از نقطه‌ی اصلی، این مقدار را به توان ۲ می‌رسانیم (نسبت به قدر مطلق کار ساده‌تری است) که به ما کمک می‌کند که مقادیر منفی نداشته باشیم (البته این کار یک ایرادی هم دارد که باعث می‌شود که تاثیر داده‌های پرت کمی بیشتر هم شوند) و در نهایت هم میانگین این مقادیر را حساب می‌کنیم تا نتیجه‌ی مناسب‌تری داشته باشیم. همه‌ی این کارها در روند اندازه‌گیری خطا را به این دلیل انجام می‌دهیم تا بتوانیم مدل دقیق‌تری را بسازیم و آن را بهبود ببخشیم.

ج)

$$X^T X\beta = X^T y \rightarrow \beta = (X^T X)^{-1} X^T y$$

$$X = \begin{bmatrix} 1 & 1.58 \\ 1 & 1.6 \\ 1 & 1.62 \\ 1 & 1.65 \\ 1 & 1.68 \\ 1 & 1.7 \\ 1 & 1.74 \\ 1 & 1.75 \\ 1 & 1.77 \\ 1 & 1.8 \end{bmatrix} , \quad y = \begin{bmatrix} 57.5 \\ 58.2 \\ 59.5 \\ 62.1 \\ 63.4 \\ 64.5 \\ 66.2 \\ 67.7 \\ 69.4 \\ 71.3 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 10 & 16.89 \\ 16.89 & 28.57 \end{bmatrix} , \quad X^T y = \begin{bmatrix} 639.8 \\ 1083.82 \end{bmatrix} \rightarrow \beta = \begin{bmatrix} -40.91 \\ 62.10 \end{bmatrix}$$

$$y = 62.1x - 40.91$$

سوال ۶)

الف)

$$Entropy \rightarrow H(X) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

احتمال (درصد)	طبقه اقتصادی (X)
۳۵	ضعیف
۲۷.۵	متوسط
۳۷.۵	مرفه

حزب مورد علاقه (Y)	احتمال (درصد)
دموکرات	۵۰
جمهوری خواه	۵۰

$$H(X) = 1.57$$

$$H(Y) = 1$$

ب)

$$Mutual\ Information \rightarrow I(X,Y) = H(X) + H(Y) - H(X,Y)$$

$$H(X,Y) = -\sum_i \sum_j p_{ij} \log_2 p_{ij}$$

حزب مورد علاقه	طبقه اقتصادی	احتمال (درصد)
دموکرات	ضعیف	۱۰
دموکرات	متوسط	۱۷.۵
دموکرات	مرفه	۲۲.۵
جمهوری خواه	ضعیف	۲۵
جمهوری خواه	متوسط	۱۰
جمهوری خواه	مرفه	۱۵

$$H(X,Y) = -\begin{pmatrix}(0.1 \times 3.321) + (0.175 \times 2.514) + (0.225 \times 2.152) \\ +(0.25 \times 2) + (0.1 \times 3.321) + (0.15 \times 2.736)\end{pmatrix} = 2.49$$

$$I(X,Y) = 1.57 + 1 - 2.49 = 0.08$$

ج) دو پیشامد وقتی مستقل هستند که $P(AB) = P(A)P(B)$ که در این سوال همچین حالتی برقرار نیست و این دو متغیر مستقل از هم نیستند.

وقتی دو متغیر از هم مستقل باشند، Mutual Information برابر صفر می‌شود و ما اطلاعات متقابلی از این متغیرها نسبت به هم نداریم و اگر مستقل نباشند، این حالت پیش نمی‌آید.

سوال ۷)

الف) برخی از روش‌های پیش‌پردازش داده عبارتند از:

- Aggregation یا یکپارچه‌سازی: در این روش چند ویژگی را با هم ترکیب می‌کنیم که استفاده‌های مختلفی دارد. ممکن است یک ویژگی کلی‌تر ایجاد کنیم که نتیجه‌ی مورد نیاز ما را بهتر بتواند نشان دهد. به عنوان مثال می‌توانیم شهرها را به استان‌ها یا مناطق تبدیل کنیم. این یکپارچه‌سازی ممکن است حتی صرفا برای کاهش ویژگی‌ها انجام شود.
- Sampling یا نمونه‌برداری: در این روش بخشی از داده‌ها را برای تحلیل و بررسی انتخاب می‌کنیم زیرا ممکن است که پردازش تمامی داده‌ها زمان زیادی از ما بگیرد. این نمونه باید نمایان‌گر ویژگی‌های کل داده‌های ما باشد تا بتواند نتیجه‌ی قابل اعتمادی به ما بدهد.
- Dimensionality Reduction یا کاهش ابعاد: در این روش ویژگی‌های نامربوط به مسئله را کنار می‌گذاریم که می‌تواند به ما کمک کند نتیجه‌ی بهتری بگیریم و حتی در نمایش داده‌ها نیز موفق‌تر عمل کنیم و بتوانیم با زمان و هزینه‌ی کمتری کار را به پایان برسانیم.

- Feature Selection: این روش هم در راستای همان کاهش ابعاد است که یعنی ویژگی‌هایی که برای ما اهمیت دارد را فقط انتخاب می‌کنیم. مثلا در مسئله‌ی پیش‌بینی کردن نتیجه‌ی افراد در یک آزمون، اهمیتی ندارد که نام آن‌ها چیست و باید این ویژگی‌ها را کنار بگذاریم و انتخابشان نکنیم.
- Feature Creation یا خلق ویژگی: در این روش سعی می‌کنیم ویژگی‌هایی درست کنیم که بتوانیم از طریق آن‌ها به نتایج بهتری برسیم. مثلا می‌توانیم داده‌ها را با یک تبدیل به یک فضای دیگر نگاشت کنیم زیرا این طور به نظر می‌رسد که در آن صورت نتایج بهتری می‌گیریم یا می‌توانیم یک ویژگی را خودمان از طریق ویژگی‌های دیگر بسازیم که بیشتر به کار ما می‌آید.
- Discretization و Binarization: در این روش‌ها نیز به نوعی داده‌ها را دسته‌بندی می‌کنیم به نوعی که برای خودمان قابل استفاده‌تر باشند. مثلا قد انسان‌ها را به صورت ترتیبی و به صورت کوتاه و متوسط و بلند بیان و دسته‌بندی می‌کنیم.
- Attribute Transformation: در اینجا نیز داده‌ها را به نوعی تغییر می‌دهیم که بیشتر کارایی داشته باشند. روش‌های Normalization در این دسته قرار می‌گیرند.

ب)

در روش بیشینه‌کمینه باید با کمک فرمول زیر داده‌ها را تغییر دهیم:

$$x_i' = \frac{x_i - min}{\max - min}(new_{max} - new_{mean}) + new_{mean} = \frac{x_i - 200}{1000 - 200}(1 - 0) + 0$$

پس داده‌های جدید ما برابرند با: $0, 0.125, 0.25, 0.5, 1$

در روش z-score هم باید از فرمول زیر استفاده کنیم:

$$x_i' = \frac{x_i - \bar{A}}{\sigma_A} = \frac{x_i - 500}{316.22}$$

پس داده‌های جدید ما برابرند با: $-0.94, , -0.63, -0.31, 0.31, 1.58$

سوال ۸)

الف)

$$f(B) = \|X\beta - y\|_2^2 + \alpha\|\beta\|_2^2 = (X\beta - y)^T(X\beta - y) + \alpha\beta^T\beta$$

$$\frac{\partial f(B)}{\partial \beta} = 2\,X^T(X\beta - y) + 2\alpha\beta$$

برای اینکه مقدار کمینه را بیابیم باید این گرادیان را برابر صفر قرار دهیم تا معادله نرمال به دست آید؛

$$2\,X^T(X\beta - y) + 2\alpha\beta = 0 \;\rightarrow\; \beta = (X^TX + \alpha I)^{-1}\,X^Ty$$

ب)

$$\beta_{k+1} = \beta_k - A\,\frac{\partial f(B)}{\partial \beta} = \beta_k - A(2\,X^T(X\beta_k - y) + 2\alpha\beta_k)$$

ج) طبق تعریف تابع محدب داریم:

$$0 \le \theta \le 1 \;\rightarrow f(\theta x + (1-\theta)y) \le\; \theta f(x) + (1-\theta)f(y)$$

حالا ما یک تابع g برابر $g(\beta) = \alpha\|\beta\|_2^2$ را طبق صورت سوال تعریف می‌کنیم و تعریف تابع محدب را روی آن پیاده می‌کنیم:

$$\alpha\left\|\theta\beta_x + (1-\theta)\beta_y\right\|_2^2 \le \theta\alpha\|\beta_x\|_2^2 + (1-\theta)\alpha\left\|\beta_y\right\|_2^2$$

$$\alpha(\left\|\theta\beta_x + (1-\theta)\beta_y\right\|_2^2) \le \alpha(\theta\|\beta_x\|_2^2 + (1-\theta)\left\|\beta_y\right\|_2^2)$$

می‌بینیم که رابطه‌ی بالا برای αهای نامنفی برقرار است پس ترم منظم‌ساز یک ترم محدب است.

بخش پیاده‌سازی

قسمت اول)

۱) اطلاعات دیتاست covid را در زیر مشاهده می‌کنیم:

	id	sex	birth_year	country	region	infection_reason	infected_by	confirmed_date	state
0	1	female	1984.0	China	filtered at airport	visit to Wuhan	NaN	1/20/2020	released
1	2	male	1964.0	Korea	filtered at airport	visit to Wuhan	NaN	1/24/2020	released
2	3	male	1966.0	Korea	capital area	visit to Wuhan	NaN	1/26/2020	released
3	4	male	1964.0	Korea	capital area	visit to Wuhan	NaN	1/27/2020	released
4	5	male	1987.0	Korea	capital area	visit to Wuhan	NaN	1/30/2020	released
...	...	...	...	...	...	...	...	...	...
171	172	female	1997.0	Korea	Gyeongsangbuk-do	NaN	NaN	2/24/2020	isolated
172	173	male	1949.0	Korea	Daegu	NaN	NaN	2/24/2020	deceased
173	174	female	1958.0	Korea	Gyeongsangbuk-do	NaN	NaN	2/24/2020	isolated
174	175	male	1997.0	Korea	Gyeongsangbuk-do	NaN	NaN	2/24/2020	isolated
175	176	female	1950.0	Korea	capital area	NaN	NaN	2/24/2020	isolated

176 rows × 9 columns

۲) همان‌طور که در جدول بالا نیز می‌بینیم این دیتاست دارای ۱۷۶ نمونه است که هر کدام از این ردیف(نمونه)ها اطلاعات یکی از مبتلایان به بیماری کوید-۱۹ است که ویژگی‌های مختلف این بیمار نظیر جنسیت، تاریخ تولد، کشور، منطقه، علت گرفتار شدن به این بیماری، عامل انتقال، تاریخ تایید بیماری و وضعیت بیمار را شامل می‌شود.

۳) مقادیر میانگین، بیشینه و انحراف از معیار ستون سال تولد برابرند با:

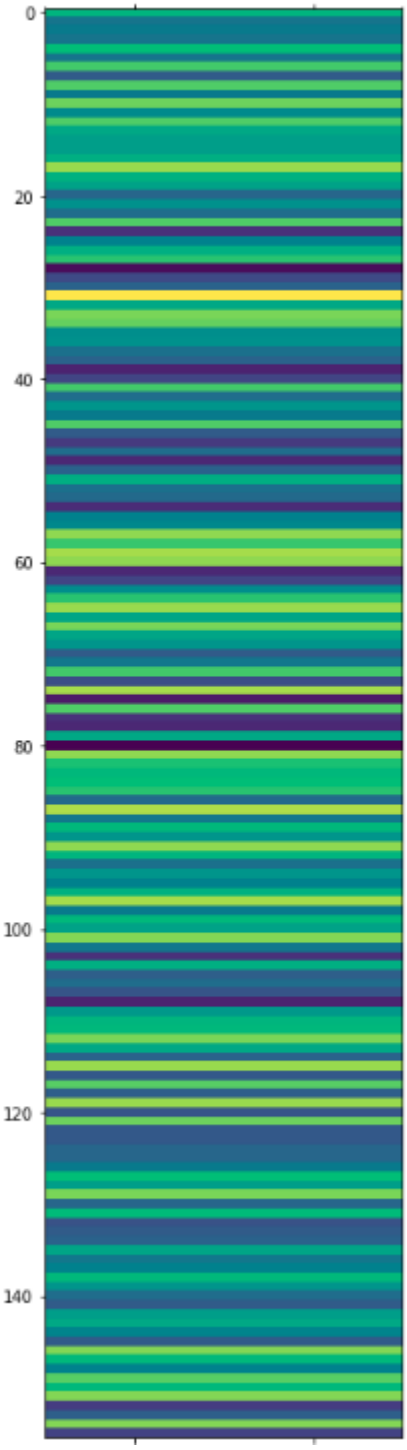
```
mean:  1973.3855421686746
max:  2009.0
std:  17.032824869574775
```

۴) تعداد سطرهایی که در هر یک از ویژگی‌های این دیتاست که مقدار ندارند به صورت زیر است:

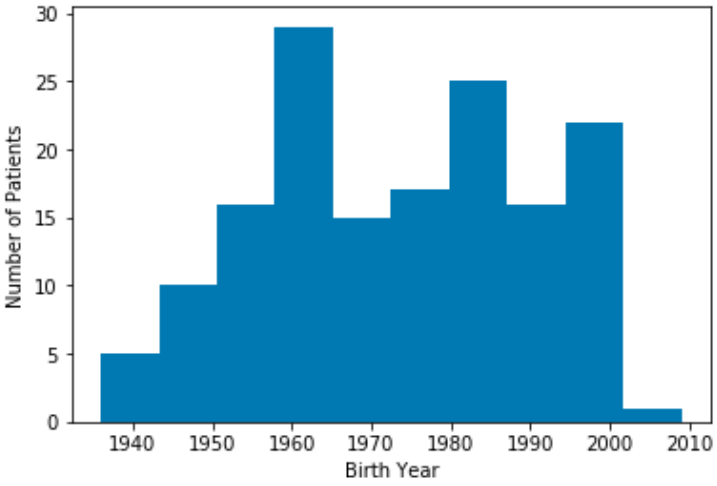
```
id                   0
sex                  0
birth_year          10
country              0
region              10
infection_reason    81
infected_by        134
confirmed_date       0
state                0
dtype: int64
```

چون برای اکثر داده‌ها، ویژگی‌های infection_reason, infected_by مقدار ندارند پس این ۲ ستون را به کلی حذف می‌کنیم و سپس سطرهایی از داده که در آن‌ها birth_year, region مقدار ندارند را حذف می‌کنیم تا همه‌ی سطرها دارای مقدار مشخص برای ویژگی‌هایشان باشند.

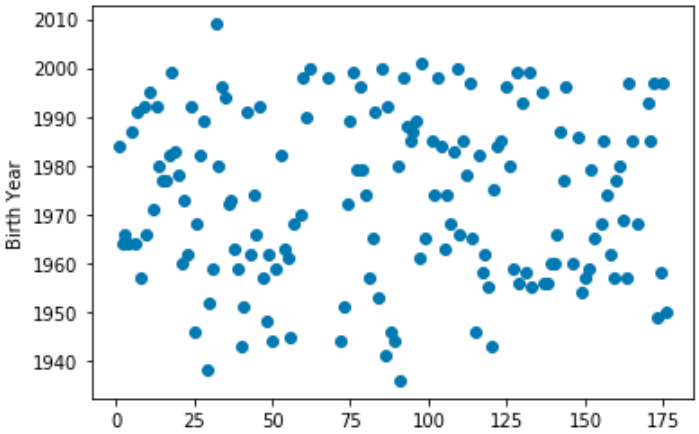
۵) نمودارهای خواسته شده را در زیر می‌بینیم:



Matrix Plot



Histogram



Scatter Plot

۶) بله به نظر چند داده‌ی outlier در ابتدا و انتهای بازه‌ی birth_year داریم. مثلا یکی از داده‌ها به تنهایی با چند سال فاصله، متولد سال ۲۰۰۹ است. چون این داده‌ها تعداد کمی هستند،برای اینکه روی مدل ما تاثیر نذارند می‌توانیم آن‌ها را حذف کنیم.

قسمت دوم)

ویژگی‌های زیر از دیتاست انتخاب شدند:

'Medu', 'Fedu', 'internet', 'schoolsup', 'studytime', famrel', 'freetime', 'Dalc', 'Walc', 'health', 'absences', 'G1', 'G2

برخی از ویژگی‌هایی که قابلیت تبدیل به حالت عددی داشتند به شکل عددی و باینری در آمدند و سپس ۸۰ درصد از داده‌ها برای یادگیری انتخاب شدند.

پس از یادگیری و اجرای مدل روی داده‌ها تست، نتایج زیر به دست آمدند:

```
R2 score:  0.8111764734264986
MSE :  3.886841982349717
```

	Test	Predicted
86	6	6.289907
165	12	11.552801
105	11	11.256873
125	12	13.668582
177	6	4.115603
...	...	...
269	0	-1.177156
59	16	16.309321
325	11	10.745684
39	13	14.140019
261	8	6.921411

79 rows × 2 columns

```
regression.score(X_test, Y_test)
```

```
0.8111764734264986
```