مهدی فیروزبخت

400131027

تمرین سری دوم – یادگیری ماشین

1)

الف)

در این سوال برای استفاده از LOOCV باید هر داده را یک بار به عنوان تست قرار داده و باقی را برای آموزش در نظر بگیریم. اگر مقدار k برابر تمام داده ها باشد تمام مقادیر تست منفی در نظر گرفته میشود که مقدار خطای 0.4 دارد.

اگر مقدار k = 1 در نظر بگیریم داده های سمت راست که همه منفی هستند و نزدیک ترین داده به آنها نیز منفی است همگی در حالت تست مقدار منفی میگیرند اما در مورد داده های سمت چپ ، اگر داده منفی تست باشد طبق آموزش مقدار آن مثبت در نظر گرفته شده و اگر داده های مثبت تست باشند ، نزدیک ترین مقدار به آنها منفی است پس مقدار آن ها نیز منفی میشود پس نیمی از حالات خطا دارند. مقدار خطای 0.5 را دارد.

اگر k = 3 باشد در این صورت داده های سمت راست درست تخمین زده میشوند اما در مورد داده های سمت چپ، مقادیر مثبت به علت اینکه در تمام حالات نزدیک 2 مثبت و 1 منفی هستند مقدار مثبت میگیرند و درست هستند اما در مورد داده منفی ، مقدار آن مثبت در نظر گرفته میشود. پس مقدار خطا برای این حالت 0.1 است.

از حالت k = 3 تا k = 6 مقدار برای مقادیر مثبت درست محاسبه میشود و برای مقدار منفی در سمت چپ تا k = 9 مقدار خطا در نظر گرفته میشود. پس از k = 3 تا k = 6 مقدار خطا ثابت است و برابر 0.1 است و بعد از آن به سمت 0.4 زیاد میشود.

اما از میان این k ها ، مقدار k = 3 بهترین k است. زیرا با افزایش k به سمت بایاس بالاتر میرویم و پیچیدگی نیست بیشتر میشود زیرا به داده ها با فواصل بیشتر میرویم. پس بهترین k برای این سوال مقدار k = 3 است.

ب)

برای یافتن بهترین مقدار k میتوانیم از الگوریتم Grid search استفاده کرد. در روش جستجوی شبکه، شبکه ای از مقادیر ممکن برای هایپرپارامترها ایجاد می کنیم. هر تکرار ترکیبی از فراپارامترها را به ترتیب خاصی امتحان می کند. این مدل را بر روی هر ترکیبی از ابرپارامترهای ممکن برازش می‌دهد و عملکرد مدل را ثبت می‌کند. در نهایت، بهترین مدل را با بهترین هایپرپارامترها برمی گرداند.

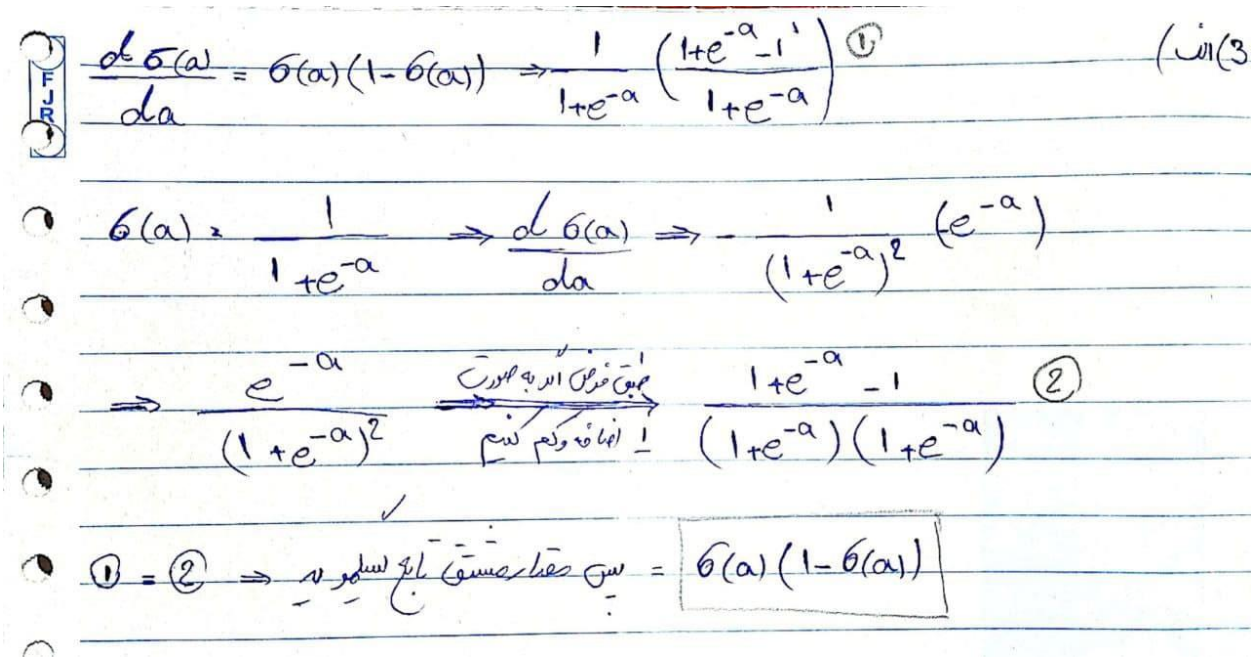
2)

KNN یک الگوریتم تمایز است زیرا احتمال مشروط یک نمونه متعلق به یک کلاس معین را مدل می کند.

درخت های تصمیم متمایز هستند زیرا مرزهای صریح بین کلاس ها را یاد می گیرند.

3)

الف)



ب)

/

① مقدار درست نمایی احتمال $\Rightarrow \prod P(y=1|x)^{y_i} \times P(y=0|x)^{1-y_i}$ ① (ب(3

$P(y=1|x) = \sigma(w^T x)$ ② $\xrightarrow{1,2}$ $likelihood = \prod \sigma(w^T x)^{y_i} \times (1-\sigma(w^T x))^{1-y_i}$

$\xrightarrow{-\log}$ $-\log(likelihood) = -\sum \log(w^T x)^{y_i} - \sum \log(1-\sigma(w^T x))^{1-y_i}$

$\Rightarrow -\sum y_i \log(w^T x) - \sum (1-y_i) \log(1-\sigma(w^T x))$ ③

$\sigma(w^T x) = \frac{1}{1+e^{-w^T x}}$, $1-\sigma(w^T x) = \frac{e^{-w^T x}}{1+e^{-w^T x}}$ ④

③,④ $\Rightarrow$ $\boxed{\sum y_i \log(1+e^{-w^T x}) - \sum (1-y_i) \log\left(\frac{e^{-w^T x}}{1+e^{-w^T x}}\right)}$

$\Rightarrow \sum w^T x_i + y_i w^T x_i + \log(1+e^{-w^T x})$

(2)

likelihood $= \sum y_i \log\left(1+e^{-\omega^T x}\right) - \sum (1-y_i) \log\left(\frac{e^{-\omega^T x}}{1+e^{-\omega^T x}}\right)$ (3

① ②

$\Rightarrow \sum y_i \log\left(1+e^{-\omega^T x}\right) - \sum (1-y_i)\left(\log e^{-\omega^T x} - \log 1+e^{-\omega^T x}\right)$

② $\Rightarrow +\sum (1+y_i)\left(-\omega^T x - \log 1+e^{-\omega^T x}\right) \Rightarrow \sum (+\omega^T x) + y(-\omega^T x) + \log(1+e^{-\omega^T x}) + y \log(1+e^{-\omega^T x})$

①,② $\Rightarrow \sum \omega^T x - y\omega^T x + \log\left(1+e^{-\omega^T x}\right) = -\log(\text{likelihood}) = LH_{new}$

$\frac{\partial LH_{new}}{\partial \omega} \Rightarrow \frac{\partial\left[\sum \omega^T x - y\omega^T x + \log\left(1+e^{-\omega^T x}\right)\right]}{\partial \omega} \Rightarrow \sum x_i - y_i x_i + \frac{-x_i e^{-\omega^T x_i}}{1+e^{-\omega^T x_i}}$

$\Rightarrow -\sum \left(y_i - 1 + \frac{e^{-\omega^T x_i}}{1+e^{-\omega^T x_i}}\right) x_i = -\sum \left(y_i + \frac{1+e^{-\omega^T x_i} - e^{-\omega^T x_i}}{1+e^{-\omega^T x_i}}\right) x_i$

$\Rightarrow -\sum \left(y_i - \frac{1}{1+e^{-\omega^T x}}\right) x_i \xrightarrow{③} \boxed{\sum \left(y_i - \hat{y}_i\right) x_i}$

$\sigma(\omega^T x) = \frac{1}{1+e^{-\omega^T x}} \xrightarrow{③} \hat{y}_i$

د)

3)ج) تابع سیگموید: $\frac{1}{1+e^{-wx}}$

در این تابع پارامتری به پیچیدگی تأثیر می‌گذارد پارامتر w است. هرچه مقدار w بیشتر شود، تابع سیگموید شیب‌تر به سمت بیش برازش می‌رود. هرچه مقدار w کمتر باشد حالت تابع به صورت (⌐) خواهد بود که اجازه آموزش می‌دهد اما اگر مقدار w بیشتر شود این منحنی به سمت حالت (⌐¬) می‌رود که بیش برازش زیادی دارد.

4) الف)

(الف.4

$P(buy=yes)=\frac{9}{14}$ $\quad P(buy=No)=\frac{5}{14}$

$X_1 \Rightarrow$

$P(youth\mid yes)=\frac{2}{9}$	$P(youth\mid No)=\frac{3}{5}$
$P(high\mid yes)=\frac{2}{9}$	$P(high\mid No)=\frac{2}{5}$
$P(yes\mid yes)=\frac{6}{9}$	$P(yes\mid No)=\frac{1}{5}$
$P(fair\mid yes)=\frac{6}{9}$	$P(fair\mid No)=\frac{2}{5}$

$(\frac{2}{9})^2\times(\frac{6}{9})^2\times\frac{9}{14}=0.0141$ ✓ $\quad \frac{3}{5}\times\frac{4}{25}\times\frac{1}{5}\times\frac{5}{14}=0.0115$

$X_2 \Rightarrow$

$P(senior\mid yes)=\frac{3}{9}$	$P(senior\mid No)=\frac{2}{5}$
$P(low\mid yes)=\frac{3}{9}$	$P(low\mid No)=\frac{1}{5}$
$P(No\mid yes)=\frac{3}{9}$	$P(No\mid No)=\frac{4}{5}$
$P(ex\mid yes)=\frac{3}{9}$	$P(ex\mid No)=\frac{3}{5}$

$(\frac{3}{9})^4\times\frac{9}{14}=0.079$ $\quad \frac{5}{14}\times\frac{2}{5}\times\frac{1}{5}\times\frac{4}{5}\times\frac{3}{5}=0.0137$ ✓

$X_3 \Rightarrow$

$P(middle\mid yes)=\frac{4}{9}$	$P(middle\mid No)=\frac{0}{5}$
$P(medium\mid yes)=\frac{4}{9}$	$P(medium\mid No)=\frac{2}{5}$
$P(No\mid yes)=\frac{3}{9}$	$P(No\mid No)=\frac{4}{5}$
$P(fair\mid yes)=\frac{6}{9}$	$P(fair\mid No)=\frac{2}{5}$

$P(middle\mid No)$ باشد، از Smoothed likelihood استفاده می‌کنیم تا Zero Prob رخ ندهد:

$\frac{0+L}{5+L\times K} \Rightarrow L=1, K=4 \Rightarrow \frac{1}{9}$: استفاده می‌کنیم

$(\frac{4}{9})^2\times\frac{3}{9}\times\frac{6}{9}\times\frac{9}{14}=0.281$ ✓ $\quad \frac{1}{9}\times(\frac{2}{5})^2\times\frac{4}{5}\times\frac{5}{14}=0.005$

ب)

۴.ب) $S(age) \Rightarrow S[youth] = [2_+, 3_-]$, $S[middle] = [4_+, 0_-]$, $S[senior] = [3_+, 2_-]$

$\Rightarrow E[youth] = -(\frac{2}{5}\log\frac{2}{5} + \frac{3}{5}\log\frac{3}{5}) = 0.970$ $E[middle] = 0$, $E[senior] = 0.970$

$E(All) = 0.940$ $gain(age) = 0.940 - \frac{5}{14}\times(0.970) - \frac{5}{14}\times(0.970) = \boxed{0.226}$

$S[income] \Rightarrow S[high] = [2_+, 2_-]$, $S[medium] = [4_+, 2_-]$, $S[low] = [3_+, 1_-]$

$E[high] = 1$, $E[medium] = 0.917$, $E[low] = 0.811$

$gain[income] = 0.940 - (0.917)\times\frac{6}{14} - (0.811)\times\frac{4}{14} = \boxed{0.031}$

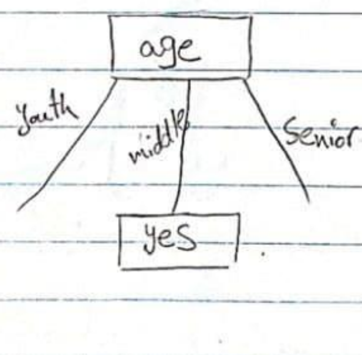
$S[stu] \Rightarrow S[yes] = [6_+, 1_-]$, $S[No] = [3_+, 4_-]$ $\Rightarrow E[yes] = 0.591$, $E[No] = 0.984$

$gain[stu] = 0.940 - \frac{7}{14}(0.591) - \frac{7}{14}(0.984) = \boxed{0.153}$

$S[credit] \Rightarrow S[fair] = [6_+, 2_-]$, $S[ex] = [3_+, 3_-]$ $\Rightarrow E[fair] = 0.811$, $E[ex] = 1$

$gain[credit] = 0.940 - \frac{8}{14}(0.811) - \frac{6}{14} = \boxed{0.048}$

ریشه age قرار میگیرد:



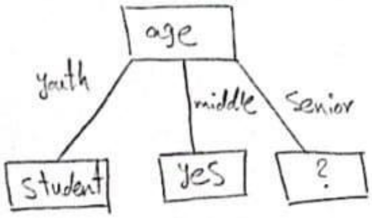
با در نظر گرفتن age=youth

$E(All) = 0.970$

$S[income] \Rightarrow E[high] = 0, E[medium] = 1, E[low] = 0 \Rightarrow gain[income] = .57$

$S[stu] \Rightarrow E[yes] = 0, E[No] = 0 \Rightarrow gain[stu] = 0.970$ ✓

$S[credit] \Rightarrow E[fair] = 0.917, E[en] = 1 \Rightarrow gain[credit] = 0.0192$



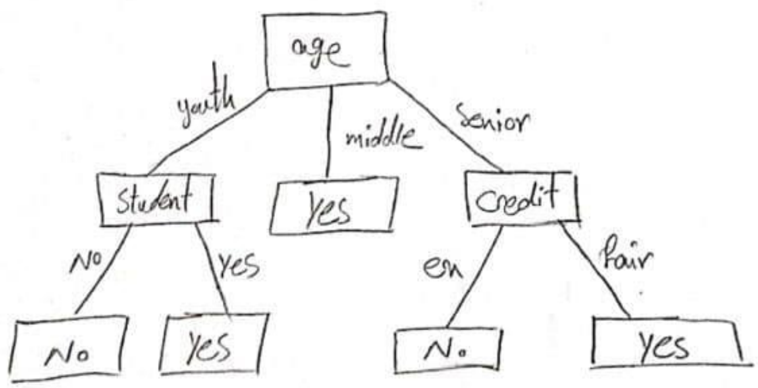
در age=Senior

$E(All) = 0.970$

$S[income] \Rightarrow E[medium] = 0.917, E[low] = 1 \Rightarrow gain(income) = 0.020$

$S[stu] \Rightarrow E[yes] = 0.917, E[No] = 1 \Rightarrow gain(stu) = 0.020$

$S[credit] \Rightarrow E[fair] = 0, E[en] = 0 \Rightarrow gain(credit) = 0.970$



چون آنتروپی credit در راست و student در چپ برابر ۰ شد، به انتهای درخت یعنی برگ رسیده‌ایم.

پس ویژگی Income از طریق ۲ ویژگی دیگر قابل محاسبه است.

5)

الف)

```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

wage-increase-first-year <= 2.5: bad (15.27/2.27)
wage-increase-first-year > 2.5
|   statutory-holidays <= 10: bad (10.77/4.77)
|   statutory-holidays > 10: good (30.96/1.0)

Number of Leaves  :     3

Size of the tree :      5


Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          42               73.6842 %
Incorrectly Classified Instances        15               26.3158 %
Kappa statistic                          0.4415
Mean absolute error                      0.3192
Root mean squared error                  0.4669
Relative absolute error                 69.7715 %
Root relative squared error             97.7888 %
Total Number of Instances               57

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.700    0.243    0.609      0.700    0.651      0.444    0.695     0.559     bad
               0.757    0.300    0.824      0.757    0.789      0.444    0.695     0.738     good
Weighted Avg.  0.737    0.280    0.748      0.737    0.740      0.444    0.695     0.675
```

ب)

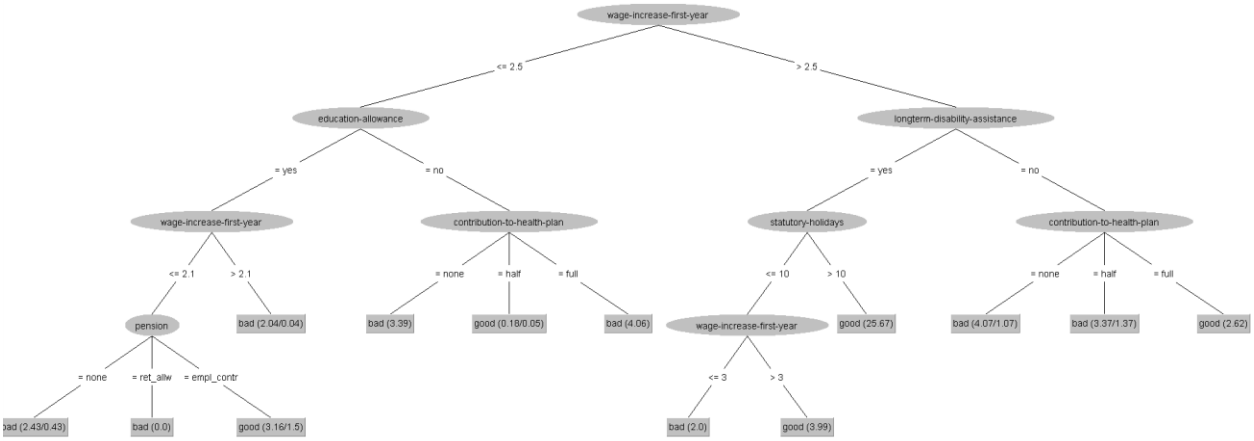
```
=== Confusion Matrix ===

  a  b   <-- classified as
 14  6 |  a = bad
  9 28 |  b = good
```

ج)

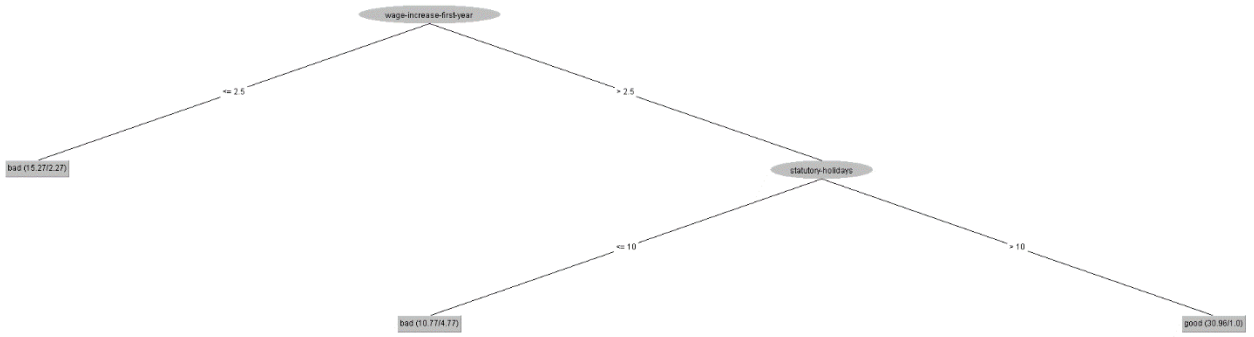
```
=== Classifier model (full training set) ===

J48 unpruned tree
------------------

wage-increase-first-year <= 2.5
|   education-allowance = yes
|   |   wage-increase-first-year <= 2.1
|   |   |   pension = none: bad (2.43/0.43)
|   |   |   pension = ret_allw: bad (0.0)
|   |   |   pension = empl_contr: good (3.16/1.5)
|   |   wage-increase-first-year > 2.1: bad (2.04/0.04)
|   education-allowance = no
|   |   contribution-to-health-plan = none: bad (3.39)
|   |   contribution-to-health-plan = half: good (0.18/0.05)
|   |   contribution-to-health-plan = full: bad (4.06)
wage-increase-first-year > 2.5
|   longterm-disability-assistance = yes
|   |   statutory-holidays <= 10
|   |   |   wage-increase-first-year <= 3: bad (2.0)
|   |   |   wage-increase-first-year > 3: good (3.99)
|   |   statutory-holidays > 10: good (25.67)
|   longterm-disability-assistance = no
|   |   contribution-to-health-plan = none: bad (4.07/1.07)
|   |   contribution-to-health-plan = half: bad (3.37/1.37)
|   |   contribution-to-health-plan = full: good (2.62)

Number of Leaves  :     13

Size of the tree :      22


Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances          45               78.9474 %
Incorrectly Classified Instances        12               21.0526 %
Kappa statistic                          0.5378
Mean absolute error                      0.2677
Root mean squared error                  0.432
Relative absolute error                 58.5226 %
Root relative squared error             90.4708 %
Total Number of Instances               57
```

د)

بدون هرس :



با هرس :



در این دیتاست ، در حالت با هرس عمق درخت 2 است در صورتی که بدون هرس عمق 4 است ، همچنین تعداد ویژگی ها در با هرس برابر 2 است در صورتی که برای بدون هرس مقدار آن 6 است که نشان دهنده پیچیدگی است. برای داده های کمتر از 2.5 برای حالت بدون هرس ممکن است حالت خوبی هم باشد در صورتی که در حالت با هرس برای داده ها کمتر از 2.5 حالت بد است.

پیاده سازی:

1)

2)

الف)

برای حل این مشکل ما میتوانیم به جای الگوریتم ID3 از الگوریتم C4.5 استفاده کنیم که در این الگوریتم مقادیر نامعلوم قابل حل هستند.

الگوریتم C4.5 با برگرداندن توزیع احتمال برچسب ها در زیر شاخه مشخصه ای که مقدار آن وجود ندارد، با مقادیر گمشده سروکار دارد. فرض کنید نمونه‌ای در داده‌های آزمایشی خود داشتیم که آب و هوا را آفتابی نشان می‌داد اما مقداری برای ویژگی رطوبت نداشت. همچنین، فرض کنید که داده‌های آموزشی ما دارای 2 نمونه بود که آب و هوا آن آفتابی، رطوبت زیر 75 و برچسب Play بود. علاوه بر این، فرض کنید داده‌های آموزشی 3 مورد داشتند که در آن آب و هوا آفتابی، رطوبت بالای 75 بود، و دارای برچسب Don't Play بود. بنابراین برای نمونه آزمایشی با ویژگی رطوبت از دست رفته، الگوریتم C4.5 توزیع احتمال [0.4، 0.6] مربوط به [Play, Don't Play] را برمی گرداند.

ج)

داده های تست شده و مقادیر محاسبه شده برای در فایل Final.csv در پوشه قرار داده شده است.

3)

الف و ب)

```
For K =  1 Average of Validations  96.00000000000001
For K =  3 Average of Validations  94.66666666666666
For K =  5 Average of Validations  93.33333333333334
For K =  7 Average of Validations  93.33333333333334
For K =  9 Average of Validations  93.33333333333334
For K =  11 Average of Validations  93.33333333333334
For K =  13 Average of Validations  94.0
For K =  17 Average of Validations  92.66666666666667
For K =  21 Average of Validations  92.0
For K =  23 Average of Validations  92.66666666666667
For K =  25 Average of Validations  91.33333333333333
For K =  27 Average of Validations  90.66666666666667
For K =  29 Average of Validations  90.66666666666667
For K =  31 Average of Validations  90.0
For K =  33 Average of Validations  89.33333333333333
For K =  35 Average of Validations  87.99999999999999
For K =  37 Average of Validations  88.0
For K =  39 Average of Validations  87.33333333333334
For K =  41 Average of Validations  86.66666666666667
For K =  43 Average of Validations  87.33333333333333
For K =  45 Average of Validations  86.66666666666667
For K =  47 Average of Validations  86.66666666666667
For K =  49 Average of Validations  86.0
For K =  51 Average of Validations  86.0
```

مقدار هر k و همچنین میانگین accuracy برای cross validation هر k محاسبه شده است و طبق این مقادیر بدست آمده مقدار k = 1 بهترین حالت برای این دیتاست است.

مقادیر ماتریس درهم ریختگی برای هر داده در فایل پیوست قرار داده شده است.