HW3

DATA MINING

Aisan Aghazadeh 9331001

AUT

Computer and Information Technology

Engineering Department

سؤال ۱

در این تمرین داده‌ها تنها به صورت متن یک sms هستند و هیچ ویژگی دیگری در اختیار نیست.
پس از خواندن فایل csv. می‌بینیم که داده‌ی یادگیری ۵ ستون دارد که یک ستون آن خروجی است، یک ستون متن پیام و ۳ ستون دیگر در اکثر سطرها مقدار ندارند و با توجه به اینکه پر کردن این ستون‌ها با توجه با مقدار موجود خطای خیلی بالایی خواهد داشت این ۳ ستون را حذف می‌کنیم:

```
del (train_set['v3'])
del (train_set['v4'])
del (train_set['v5'])
```

در اکثر روش‌های یادگیری ماشین برای بالا بردن دقت خطا، داده‌ها را به دو بخش train و test تقسیم می‌کنند و یادگیری روی train انجام می‌شود و ارزیابی بر روی داده‌های test انجام می‌شود.

```
x_train, x_test, y_train, y_test = train_test_split(train_set['content'],
train_set['label'], random_state = 1)
```

در اکثر مسائل پردازش متن، متن را به صورت ترکیب کلمات می‌بینیم. در این مسائله متن را به یک bag of words تبدیل می‌کنیم.

```
count_vector = CountVectorizer()
training_data = count_vector.fit_transform(x_train)
print(training_data)
testing_data = count_vector.transform((x_test))
```

پس از این کار با استفاده از sklearn مدلی بر اساس بیز ساده می‌سازیم:

```
naive_bayes = MultinomialNB()
naive_bayes.fit(training_data, y_train)
```

در نهایت داده‌های تست را که آن‌ها هم به صورت یک bag of words درآمده است را به مدل می‌دهیم و پیش‌بینی را انجام می‌دهیم:

```
preditions = naive_bayes.predict(testing_data)
```

پس این مرحله برای ارزیابی ۳ معیار Recall، Precision،Accuracy را محاسبه می‌کنیم:

```
print(accuracy_score(y_test, preditions))
print(recall_score(y_test, preditions))
print(precision_score(y_test, preditions)
```

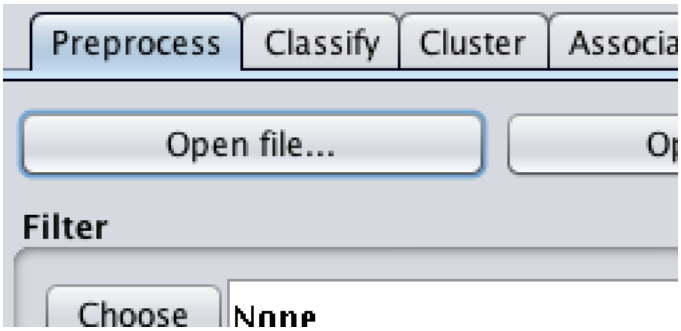
خروجی به شکل زیر می‌شود:

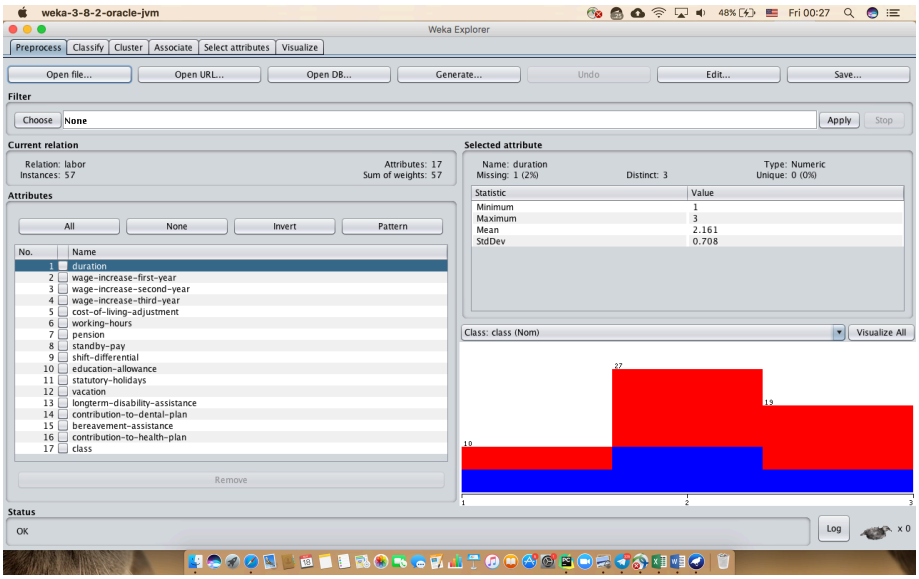
```
0.985642498205
0.933333333333
0.954545454545
```

می‌توان گفت در این حل معیار precision مقدار بیشتری دارد در نتیجه احتمال اینکه غلط باشد و بگوییم درست است کمتر است. پس اینکه خطای نوع دوم رخ دهد احتمال کمتری دارد.

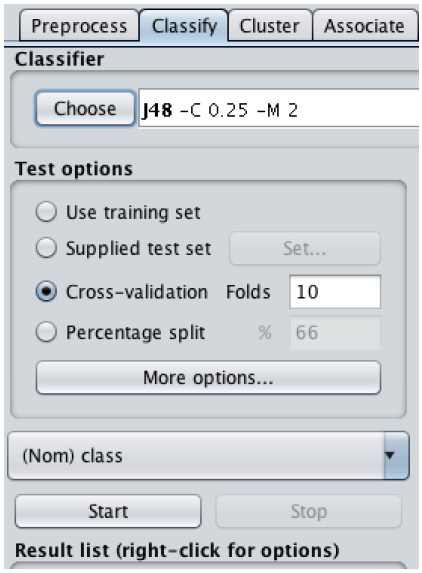
سؤال ۲

فایل را باز می‌کنیم:





وارد تب classify می‌شویم و تنظیمات را به صورت زیر تغییر داده و start را می‌زنیم:



خروجی برای مدل اول به شکل زیر می‌شود:

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     labor
Instances:    57
Attributes:   17
              duration
              wage-increase-first-year
              wage-increase-second-year
              wage-increase-third-year
              cost-of-living-adjustment
              working-hours
              pension
              standby-pay
              shift-differential
              education-allowance
              statutory-holidays
              vacation
              longterm-disability-assistance
              contribution-to-dental-plan
              bereavement-assistance
              contribution-to-health-plan
              class
Test mode:    10-fold cross-validation

=== 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	

=== Confusion Matrix ===

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

Accuracy: Correctly Classified Instances 42 73.6842 %

Recall, precision and F-measure:

```
=== 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
```

Confusion Matrix:

```
=== Confusion Matrix ===

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

precision:

$$precision_a = \frac{TP}{TP + FP} = \frac{14}{14 + 9} = 0.609$$
$$precision_b = \frac{TP}{TP + FP} = \frac{28}{6 + 28} = 0.824$$

Recall:

$$recall_a = \frac{TP}{TP + FN} = \frac{14}{14 + 6} = 0.7$$
$$recall_b = \frac{TP}{TP + FN} = \frac{28}{28 + 9} = 0.757$$

F-measure:

$$F - measure_a = \frac{2 \times recall \times precision}{recall + precision} = \frac{2 \times 0.7 \times 0.609}{0.7 + 0.609} = 0.651$$

$$F - measure_b = \frac{2 \times recall \times precision}{recall + precision} = \frac{2 \times 0.757 \times 0.824}{0.757 + 0.824} = 0.789$$

classification:

```
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)
```

output:

good: wage-increase-first-year = 3 > 2.5 and statutory-holidays = 12 > 10

همین مراحل برای مدل دوم طی می‌شود و خروجی به شکل زیر می‌شود:

```
=== Run information ===

Scheme:       weka.classifiers.trees.DecisionStump
Relation:     labor
Instances:    57
Attributes:   17
              duration
              wage-increase-first-year
              wage-increase-second-year
              wage-increase-third-year
              cost-of-living-adjustment
              working-hours
              pension
              standby-pay
              shift-differential
              education-allowance
              statutory-holidays
              vacation
              longterm-disability-assistance
              contribution-to-dental-plan
              bereavement-assistance
              contribution-to-health-plan
              class
Test mode:    10-fold cross-validation

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

Decision Stump

Classifications

pension = none : bad
pension != none : good
pension is missing : good

Class distributions
```

pension = none
bad good
1.0 0.0
pension != none
bad good
0.4375 0.5625
pension is missing
bad good
0.06666666666666667 0.9333333333333333

Time taken to build model: 0 seconds

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

Correctly Classified Instances	46	80.7018 %
Incorrectly Classified Instances	11	19.2982 %
Kappa statistic	0.5393	
Mean absolute error	0.2102	
Root mean squared error	0.3358	
Relative absolute error	45.9597 %	
Root relative squared error	70.3345 %	
Total Number of Instances	57	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.550	0.054	0.846	0.550	0.667	0.564	0.835	0.815	bad
	0.946	0.450	0.795	0.946	0.864	0.564	0.835	0.851	good
Weighted Avg.	0.807	0.311	0.813	0.807	0.795	0.564	0.835	0.838	

=== Confusion Matrix ===

a	b	<-- classified as
11	9	a = bad
2	35	b = good

Acuracy: Correctly Classified Instances 46 80.7018 %

Recall, precision and F-measure:

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.550    0.054    0.846      0.550    0.667      0.564    0.835     0.815     bad
                 0.946    0.450    0.795      0.946    0.864      0.564    0.835     0.851     good
Weighted Avg.    0.807    0.311    0.813      0.807    0.795      0.564    0.835     0.838
```

Confusion Matrix:

```
=== Confusion Matrix ===

  a  b   <-- classified as
 11  9 |  a = bad
  2 35 |  b = good
```

precision:

$$precision_a = \frac{TP}{TP + FP} = \frac{11}{11 + 2} = 0.846$$
$$precision_b = \frac{TP}{TP + FP} = \frac{35}{35 + 9} = 0.795$$

Recall:

$$recall_a = \frac{TP}{TP + FN} = \frac{11}{11 + 9} = 0.55$$
$$recall_b = \frac{TP}{TP + FN} = \frac{35}{35 + 9} = 0.946$$

F-measure:

$$F - measure_a = \frac{2 \times recall \times precision}{recall + precision} = \frac{2 \times 0.55 \times 0.846}{0.55 + 0.846} = 0.667$$

$$F - measure_b = \frac{2 \times recall \times precision}{recall + precision} = \frac{2 \times 0.946 \times 0.795}{0.946 + 0.795} = 0.864$$

classification:

```
Decision Stump

Classifications

pension = none : bad
pension != none : good
pension is missing : good
```

output:

good: pension = ret_allw != none