
بسم الله الرحمن الرحيم

دانشکده برق و کامپیوتر

پروژه داده کاوی

بخش اول: داده های بیماری قلبی و پیش بینی بیماری

بخش دوم: داده های سرطان سینه و پیش بینی سرطان

فائزه صالحی-نیلوفر اتحادی

استاد درس: سرکار خانم اسدی

1399-1398

فهرست

چکیده

در این پروژه از نرم افزار داده کاوی weka استفاده می شود. دو مجموعه داده ای که از مجموعه داده های سایت UCI انتخاب کردیم به ترتیب heart disease و Breast cancer wisconsin می باشد. روی هر کدام از این مجموعه داده ها پاکسازی ، دو روش طبقه بندی ، ارزیابی روش استفاده شده و تحلیل نتایج مفسری صورت می گیرد.

فصل اول

Heart disease

1-1- مقدمه

این مجموعه داده شامل 303 نمونه می باشد و 75 مولفه دارد ولی از این تعداد مولفه تنها 14 ویژگی اصلی وجود دارد . تمامی این ویژگی ها numeric هستند و در زیر توضیح مختصری روی آنها داده شده است :

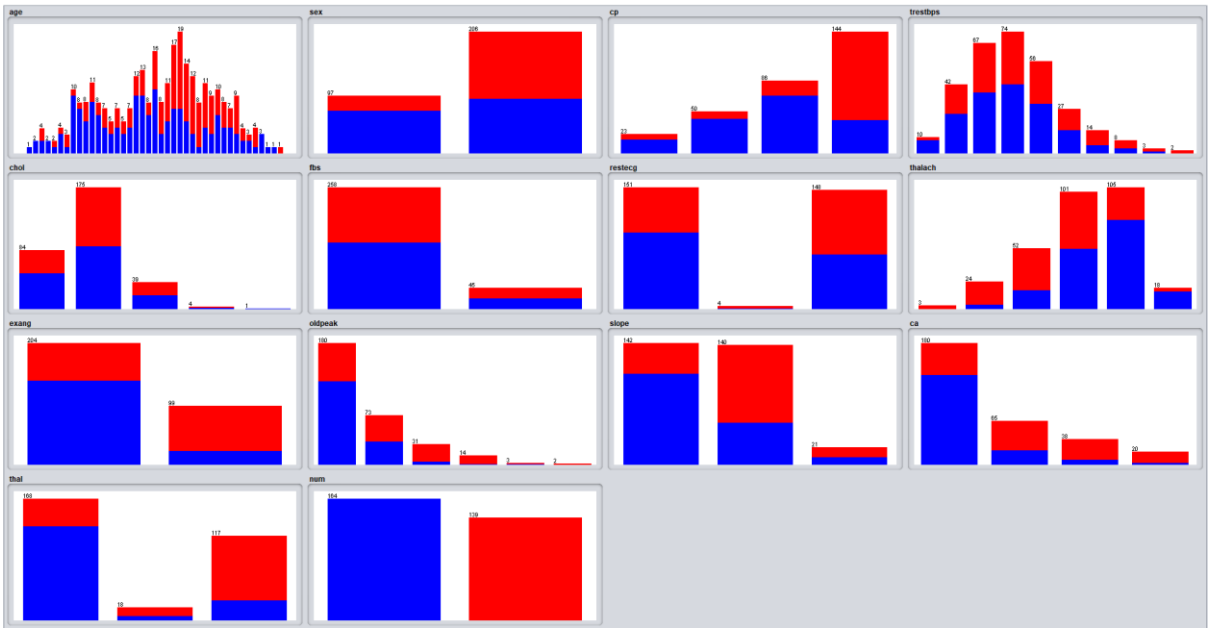
1) Age : سن هر فرد را نشان می دهد.
2) Sex: جنسیت > (male:1 , female:0)
3) Cp: درد قفسه سینه > (1= ورم گلو نرمال ، 2= ورم گلوی غیر نرمال 3=غیر وابسته به گلودرد 4=مجانبی)
4) Trestbps: فشار خون > mm Hg
5) Chol : کلسترول > mg/dl
6) Fbs : قند خون ناشتا >120 : =1 در غیر اینصورت =0
7) Restecg : نتایج الکتروکاردیوگرافی : (0= نرمال ، 1= داشتن ناهنجاری موج st-t ، 2= هیپرتروفی بطن چپ)
8) Thalach : حداکث ضربان قلب بدست آمده
9) Exang : ورزش منجر به ورم گلو (0=خیر 1=بله)
10) Oldpeak : افسردگی st ناشی از ورزش
11) Slope : شیب (مثبت=1 مسطح=2 منفی=3)
12) Ca : تعداد عروق اصلی (0-3)
13) Thal : تالاسمی (3=نرمال ، 6=نقض ثابت ، 7=نقض برگرداننده)
14) Num: تشخیص بیماری قلبی (باریک شدن قطر < 50٪ =0 ، باریک شدن قطر >50٪ =1)

کلاس ما num می باشد .

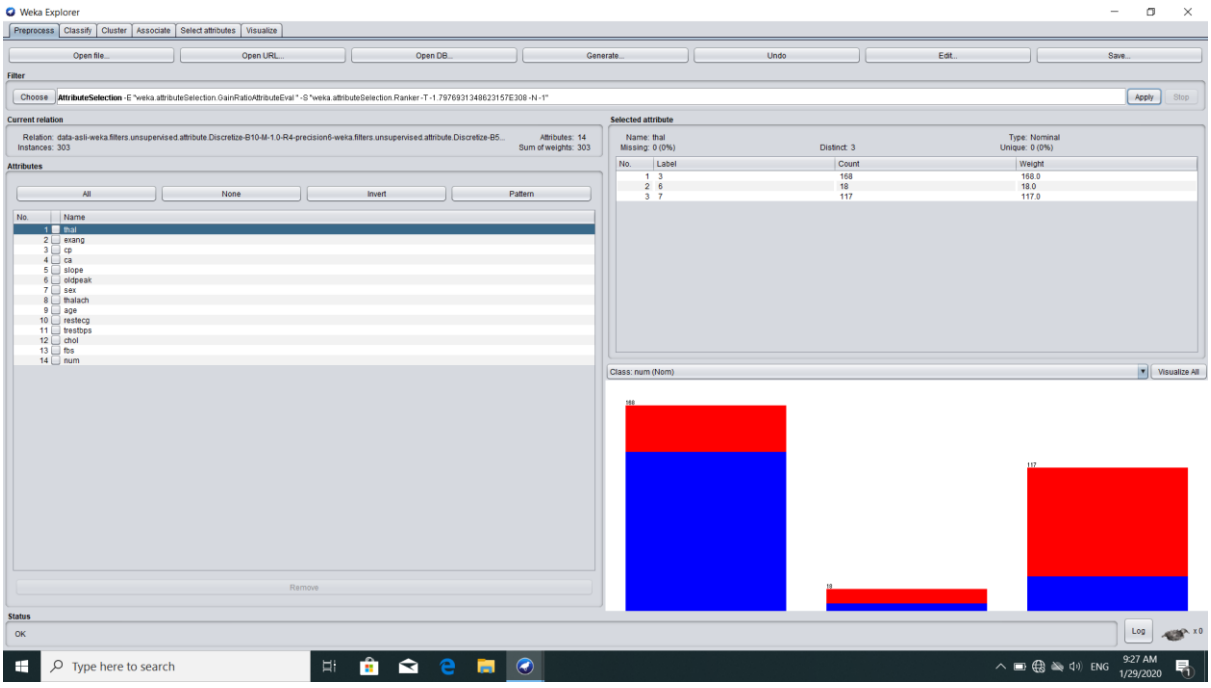
2-1- پیش پردازش داده ها

براساس مشکلاتی که در مجموعه داده مشاهده می کنیم داده را پاکسازی می کنیم .

در اولین مرحله ویژگی trestbps را با استفاده از فیلتر discretize که نمونه فیلتری است که طیف وسیعی از ویژگی های عددی را به اسمی تبدیل می کند استفاده کردم (در واقع میخواستم داده ها در 10 دسته تقسیم شوند) ، سپس در مرحله بعدی ویژگی chol را با همان فیلتر به 5 دسته تقسیم میکنیم (یعنی bin=5) ، ویژگی oldpeak و thalachرا هم با همین فیلتر به 6 دسته تقسیم کردم ، در مرحله بعدی بقیه ویژگی ها فیلتر numeric to nominal را انتخاب میکنیم . حالا همه داده ها nominal شدند ولی توجه کنید 4 نویز در ca ، 2 نویز thal ، در مجموع 6 نویز داریم از آنجا که مقادیر تهی بسیار کم است می توانیم آنها را رها کنیم یا به آنها فشار وارد کنیم ،من میانگین را به جای مقادیر تهی قرار دادم اما می توان این سطر ها را هم به طور کامل حذف کرد .(فیلتر replace missing value با میانگین جا به جا میکند. حال داده ها آماده هستند.



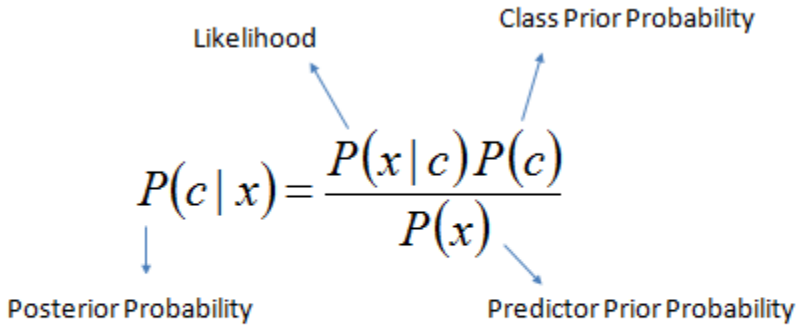
کار دیگری که در این قسمت انجام دادم فیلتر برای مشخص کردن با ارزش ترین داده تا بدترین داده است:



1-3- طبقه بندی

1-3-1-توضیحات روش های طبقه بندی

Naïve bayes: این روش به شرح زیر است :



$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

مزایای این روش پیاده سازی آسان و مقاوم بودن نسبت به نویز است. همچنین این روش در برخورد با داده های miss یا نادیده گرفتن آنها در محاسبات سایر احتمالات را محاسبه میکند و در اکثر مواقع نتایج رضایت بخشی در پی دارد. (البته این روش با فرض مستقل بودن ویژگی ها از هم کار میکند و وابستگی میان ویژگی ها را به این روش نمیشود مدل کرد).

درخت j48(c4.5):

این روش که یکی از انواع درخت های تصمیم است از معیار gain ratio جهت نرمال سازی بهره اطلاعات استفاده میکند که به روش زیر محاسبه میگردد (و البته ویژگی با حداکثر مقدار gain ratio به عنوان ویژگی تقسیم استفاده میگردد):

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{D}\right)$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)}$$

این الگوریتم مزایای بسیاری دارد و مشابه ID3 است اما بسیاری از نقاطِ ضعف الگوریتم ID3 که در C4.5 رفع شده است مانند اینکه الگوریتم C4.5 می‌تواند مقادیر گسسته یا پیوسته را در ویژگی‌ها درک کند که الگوریتمِ ID3 اولیه نمی‌توانست تفاوتِ مقادیرِ عددیِ پیوسته را درک کند. همچنین الگوریتمِ C4.5 قادر است تا مقادیری که موجود نیستند را هم تحمل کند. سومین موردی که باعث بهینه

شدن الگوریتم C4.5 نسبت به ID3 می‌شود، عملیاتِ هرس کرد جهت جلوگیری از overfitting است. الگوریتم‌هایی مانند ID3 به خاطر اینکه سعی دارند تا حد امکان شاخه و برگ داشته باشند (تا به نتیجه مورد نظر برسند) با احتمال بالاتری دارای پیچیدگی در ساخت مدل می‌شوند و این پیچیدگی در بسیاری از موارد الگوریتم را دچار overfitting و خطای بالا می‌کند. اما با عملیات هرس کردن درخت که در الگوریتم C4.5 انجام می‌شود، می‌توان مدل را به یک نقطه بهینه رساند که زیاد پیچیده نباشد (و البته زیاد هم ساده نباشد) و overfitting یا underfitting رخ ندهد.

Svm:

یک مرزی است با معیار قرار دادن بردار های آن ،بهترین دسته بندی و تفکیک بین داده ها را برای ما مشخص می کند.
برای استفاده از این الگوریتم ابتدا داده ها را پالایش میکنیم ، داده ها را عددی و نرمال کنیم و کرنل های مختلف آن را در نظر میگیریم .

Ibk

IBKیک رده بند با K همسایه نزدیک است که معیار فاصله ذکر شده را استفاده می‌کند. تعداد نزدیکترین فاصله‌ها) پیش فرض 1 (K = می‌تواند به طور صریح در ویرایشگر شیء تعیف شود. پیش‌بینی‌های متعلق به پیش از یک همسایه می‌تواند بر اساس فاصله آنها تا نمونه-های آزمایشی، وزن‌دار گردد.

1-3-1-پیاده سازی روش ها

در این پروژه 80 درصد داده ها train و 20 درصد داده ها test است .
الان گسسته سازی داده ها شده است مثلا نتیجه های زیر را می توان گرفت:
سن 58 سالگی بیشترین احتمال مبتلا شدن به این بیماری است . و بعد سن 57
زنان بیشتر در معرض خطر هستند.
حال نوبت طبقه بندی است :
الگوریتم هایی که استفاده کردیم :
Svm (smo و کرنل)
Naïve bayes
J48
Ibk

جداول مقایسه :

Algorithm	Correctly Classified Instances	Percentage of correct instance	Incorrectly Classified Instances	Percentage of incorrect instance	Time taken to build model seconds
Svm	53	٪ 86.8852	8	٪ 13.1148	0.03
Naïve bayes	52	٪ 85.2459	9	٪ 14.7541	0
J48	48	٪ 78.6885	13	٪ 21.3115	0
Ibk	50	٪ 81.9672	11	٪ 18.0328	0

جدول 1(خلاصه طبقه بندی روش ها براساس زمان و دقت)

Algorithm	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
Svm	0.7374	0.1311	0.3621	٪ 26.2719	٪ 72.1961
Naïve bayes	0.7047	0.1716	0.3441	٪ 34.3843	٪ 68.602
J48	0.5734	0.2869	0.4139	٪ 57.4752	٪ 82.5048
Ibk	0.6387	0.2144	0.4182	٪ 42.9566	٪ 83.3624

جدول 2(خلاصه خطاهای مجموعه داده)

مشاهدات :

اگر معیار دقت را در نظر بگیریم الگوریتم svm با دقت 86.8852٪ بهتر است و بعد از آن naïve bayes بهتر است . اگر معیار زمان را در نظر بگیریم بین آنهایی که زمانشان صفر است ما naïve bayes را انتخاب میکنیم .

از جدول دوم به این نتیجه میرسیم که بیشترین خطا را ibk و به دنبال آن j48 دارد.

1-4- ارزیابی

1-4-1- روش های ارزیابی طبقه بندی یا وارسی اعتبار (cross validation)

روش وارسی اعتبار یک متد توسعه‌یافته و موردپذیرش برای آنالیز صحت پیش‌بینی می‌باشد. از این روش به‌طور عمده برای زیرمجموعه‌های تصادفی و یا چندبخشی(k-fold) از مجموعه آزمون و آموزش، استفاده می‌شود. روش ذکرشده به‌عنوان یک متد نمونه‌برداری جزء که رویکردی ساده از وارسی اعتبار می‌باشد، شناخته‌شده است. در روش اعتبارسنجی k-fold، مجموعه دادها را به k بخش مجزا تقسیم می‌شوند. فرایند مدل‌سازی را برای k مرتبه تکرار می‌کنیم و در هر مرتبه k-1 بخش از داده‌ها برای پروسه آموزش استفاده می‌شود و یک بخش از داده‌ها که در فرایند آموزش، شرکت داده نشده، برای فرایند تست و اعتبارسنجی مدل پیش‌بینی کننده، مورد استفاده قرار می‌گیرد. در خاتمه از خطای پیش‌بینی محاسبه شده در هریک از k مرحله متوسط‌گیری می‌شود. مزیت استفاده از زیرمجموعه‌سازی تصادفی داده‌ها در این روش سبب می‌شود تأثیر نحوه توزیع داده‌ها برای فرایند مدل‌سازی حذف شود. واریانس نتایج حاصل از متوسط گیری برای حالتی که مقدار ، بسیار بزرگ باشد، بسیار کوچک خواهد بود.

آنالیز حساسیت

ارزیابی عملکرد الگوریتم‌های شرح داده‌شده در بالا، با استفاده از معیارهای متفاوتی بر مبنای دیدگاه حساسیت و تشخیص صورت گرفته شده‌است. حساسیت و تشخیص در آمار دو شاخص برای ارزیابی نتیجه یک دسته‌بندی دودویی (دوحالته) هستند. زمانی که بتوان داده‌ها را به دو گروه مثبت و منفی تقسیم کرد، دقت نتایج یک آزمایش که اطلاعات را به این دو دسته تقسیم می‌کند با استفاده از شاخص‌های حساسیت و ویژگی قابل اندازه‌گیری و توصیف است. معیارهای مورد استفاده در این دیدگاه به شرح زیر می‌باشند:

Tp_i مثبت صحیح از کلاس i ام (دفعاتی که هریک از شرایط سه‌گانه احتراق را به‌درستی طبقه‌بندی می‌کند).

TN_i منفی صحیح از کلاس i ام.

FP_i مثبت کاذب از کلاس i ام (دفعاتی که هریک از شرایط سه‌گانه احتراق را به‌درستی طبقه‌بندی نمی‌کند).

FN_i منفی کاذب از کلاس i ام.

Accuracy معیار صحت بیان کننده تعداد «پیش‌بینی‌های صحیح انجام شده» توسط دسته‌بند، تقسیم بر، تعداد «کل پیش‌بینی‌های انجام شده» توسط همان دسته‌بند است.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision نسبت مقداری موارد صحیح طبقه‌بندی‌شده توسط الگوریتم از یک کلاس مشخص، به کل تعداد مواردی که الگوریتم چه به‌صورت صحیح و چه به‌صورت غلط، در آن کلاس طبقه‌بندی کرده است که به‌صورت زیر محاسبه می‌شود:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall نسبت مقداری موارد صحیح طبقه‌بندی شده توسط الگوریتم از یک کلاس به تعداد موارد حاضر در کلاس مذکور که به‌صورت زیر محاسبه می‌شود:

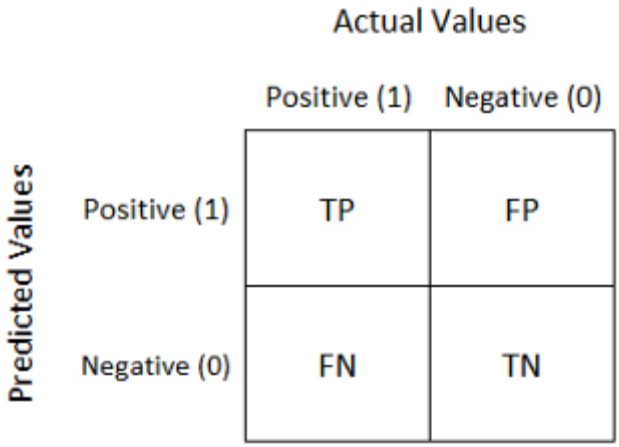
$$\text{Recall} = \frac{TP}{TP + FN}$$

F-Measure با توجه به محاسبات انجام گرفته برای معیارهای Precision و Recall، در این مرحله می‌توان مقدار کمیت وزن‌دار F-Measure را محاسبه نمود. F-Measure، پارامتر مناسبی برای ارزیابی کیفیت کلاس‌بندی می‌باشد و همچنین توصیف‌کننده میانگین وزن‌دار مابین دو کمیت Precision و Recall می‌باشد. برای یک الگوریتم کلاس‌بندی کننده در شرایط ایده‌آل، مقدار این کمیت برابر با 1 می‌باشد و در بدترین وضعیت برابر با صفر می‌باشد. این پارامتر با توجه به رابطه زیر محاسبه می‌شود:

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

1-4-2-ارزیابی مجموعه داده

توجه به جدول زیر و فرمول هایی که هست برای ارزیابی استفاده می شود:



Svm:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.903	0.167	0.848	0.903	0.875	0.739	0.868	0.816	0
	0.833	0.097	0.893	0.833	0.862	0.739	0.868	0.826	1
Weighted Avg.	0.869	0.132	0.870	0.869	0.869	0.739	0.868	0.821	

Naïve bayes:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.871	0.167	0.844	0.871	0.857	0.705	0.918	0.912	0
	0.833	0.129	0.862	0.833	0.847	0.705	0.918	0.925	1
Weighted Avg.	0.852	0.148	0.853	0.852	0.852	0.705	0.918	0.919	

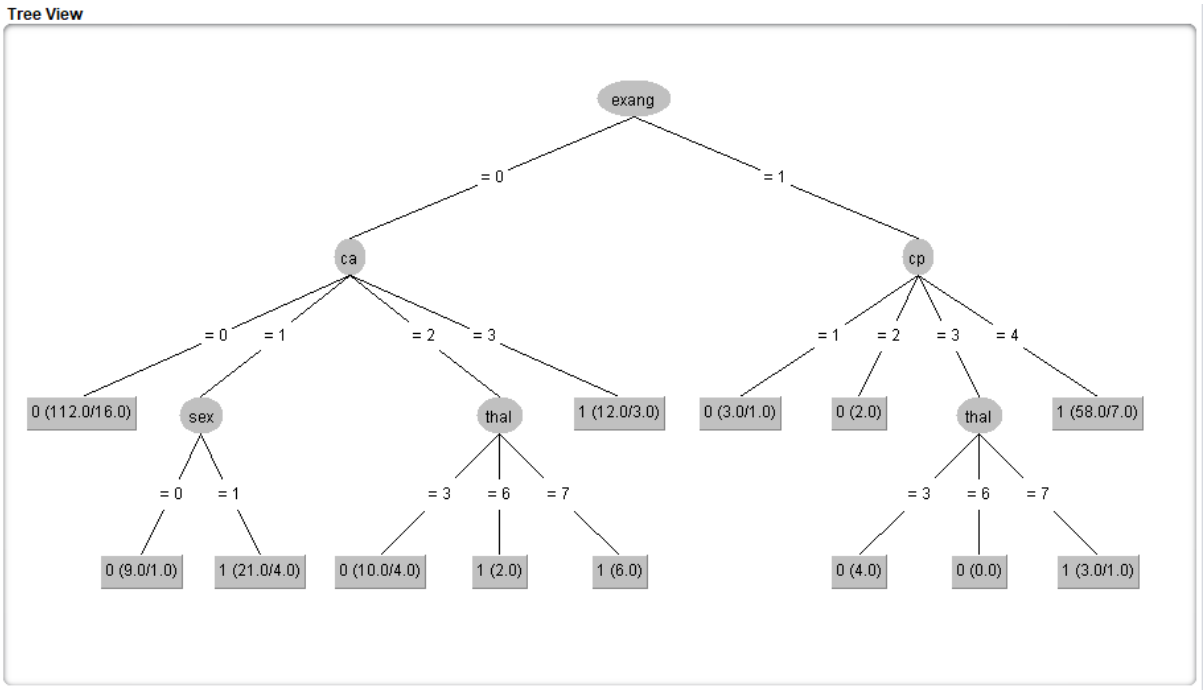
J48 :

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.806	0.233	0.781	0.806	0.794	0.574	0.812	0.747	0
	0.767	0.194	0.793	0.767	0.780	0.574	0.812	0.800	1
Weighted Avg.	0.787	0.214	0.787	0.787	0.787	0.574	0.812	0.773	

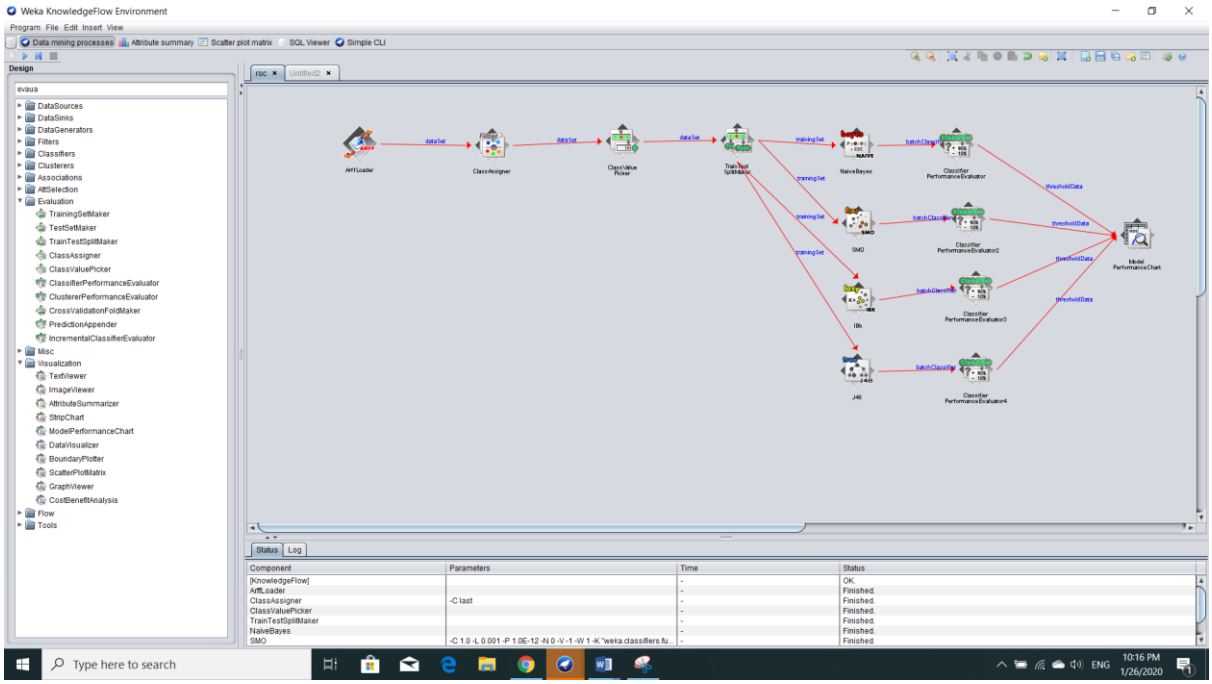
Ibk:

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.871	0.233	0.794	0.871	0.831	0.642	0.834	0.827	0
	0.767	0.129	0.852	0.767	0.807	0.642	0.831	0.776	1
Weighted Avg.	0.820	0.182	0.823	0.820	0.819	0.642	0.832	0.802	

توضیح j48 :



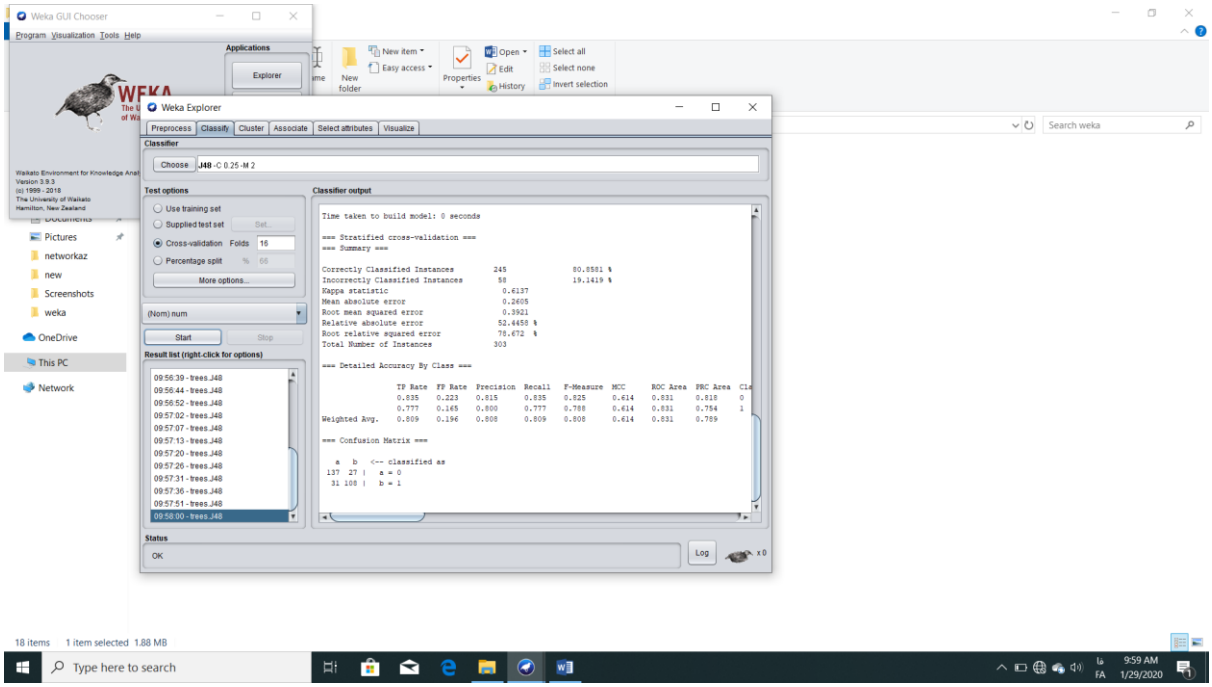
نمودار : roc



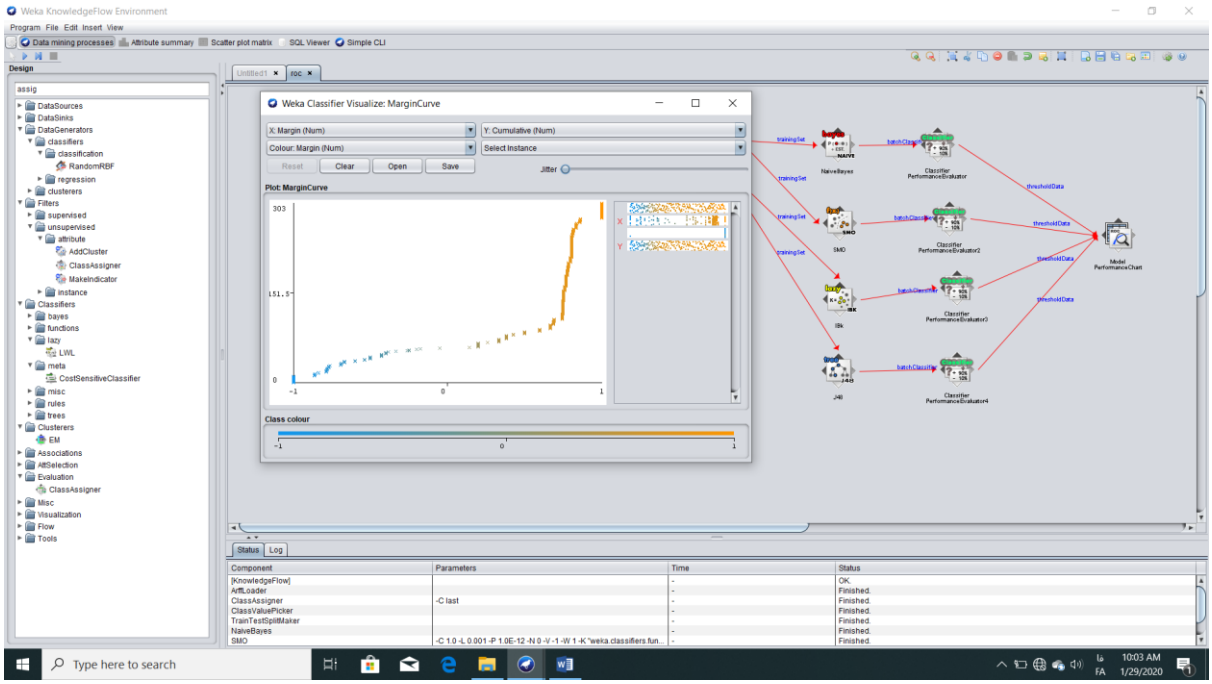
در حالتی که اول بررسی کردم از درصد استفاده کردم نه k-fold حال برای دو روش بررسی k-fold و نمودار roc را خواهیم داشت .

J48 :

K=16 : درصد درستی 80.8581 ٪ بدست آمد.

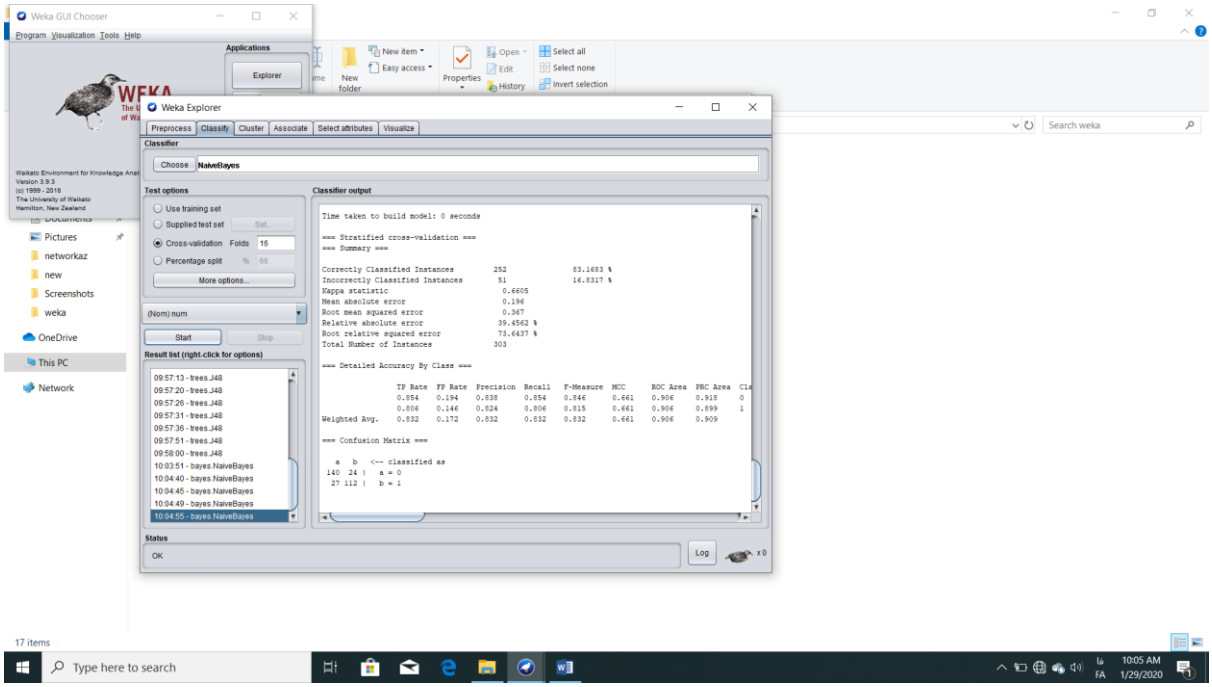


نمودار roc مربوطه :

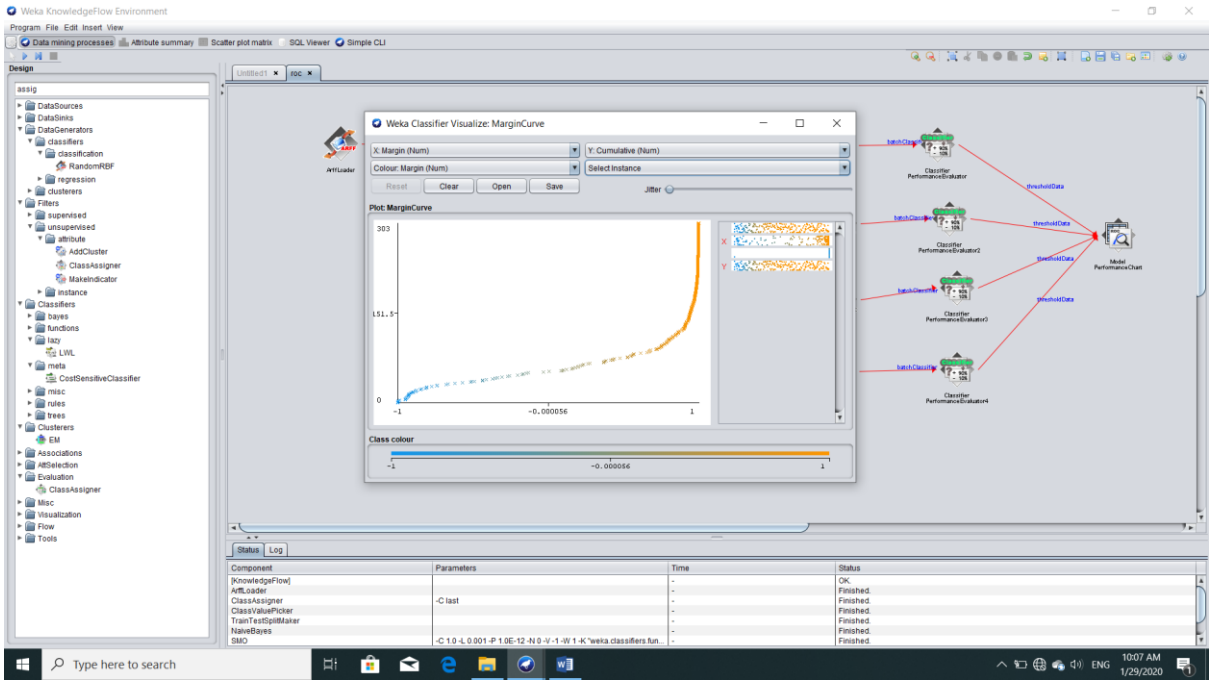


Naïve bayes :

K=16 : درصد درستی 83.1683 ٪



نمودار roc مربوطه :



در مقایسه همچنان روش naïve bayes بهتر است و مقایسه نمودار های roc هم بخاطر رشد منظم در این روش بهتر است .

فصل دوم

Breast cancer wisconsin

2-1- مقدمه

سرطان پستان از شایعترین سرطانها در میان زنان جامعه امروز میباشد. اخیرا شیوع این بیماری افزایش یافته است از آنجا که تشخیص خوشخیم یا بدخیم بودن تومور در مراحل ابتدایی این بیماری امکان درمان و عمر طوالنی مدت مبتالیان به آن را تضمین مینماید، متخصصین به دنبال روشهای بهینه جهت بهبود تشخیص این تومور می باشند.

بانک داده سرطان پستان ویسکانسین: در این مطالعه، آزمایش روی بانک اطلاعاتی ویسکانسین که از مخزن یادگیری ماشین UCIاقتباس شده است، انجام گرفته است . این بانک شامل از 698 نمونه استخراج شده ،تشکیل شده است .

ویژگی	دامنه مقادیر
ضخامت توده	10-0
یکنواختی اندازه سلول	10-0
یکنواختی شکل سلول	10-0
چسبندگی حاشیه ای	10-0
اندازه سلول مخاطی منفرد	10-0
هسته های بی تحرک	10-0
کروماتین بلاند	10-0
هسته های طبیعی	10-0
مایتروزها	10-0
کلاس	2 و 4

2-2- پیش پردازش داده ها

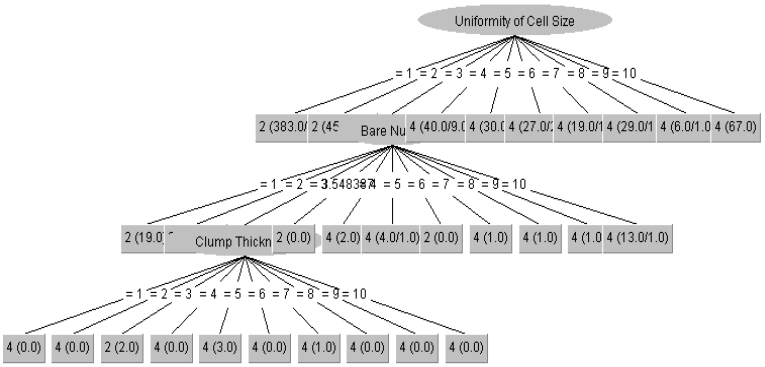
تعداد کل ویژگی های ما11 تا هستند که یکی از آنها را به جهت اینکه برای ما کاربردی ندارد حذف میکنیم(فیلد شماره نمونه ها) و نیز یکی دیگر از فیلدها برای کلاس بندی است که دارای مقادیر2(تومور خوش خیم) و 4(تومور بدخیم) است و در مورد سایر ویژگی ها که مقادیرشان در زیر نشان داده شده است دارای مقدار صحیح بین 0 تا10 میباشد. افزایش این رقم به معنای وخیمتر شدن وضعیت است، به طوری که مقدار 10 به معنای وضعیت بسیار غیرعادی است. از همه نمونه16 نمونه ناقص بوده اند که با مقادیر میانگین جایگزین شده اند (با استفاده پردازش داده ها). همچنین، از نرمالسازی وتبدیل numerictonuminal و descretize جهت پیش پردازش داده ها استفاده شده است.

از این698 نمونه 457 عدد متعلق به کلاس2 (خوش خیم) و 241عدد مربوط به کلاس4(بدخیم) هستند . از آنجایی که غده سرطانی میتواند خوش خیم (مضر نمیباشد)و بدخیم (پتانسیل مضر بودن دارد) باشد، هدف الگوریتم ارایه شده جداسازی صحیح نمونه ها به عنوان خوشخیم یا بدخیم است.

2-3- طبقه بندی

طبقه بندی به روش j48 :

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

Correctly Classified Instances         662               94.8424 %
Incorrectly Classified Instances        36                5.1576 %
Kappa statistic                          0.8864
Mean absolute error                      0.0801
Root mean squared error                  0.2108
Relative absolute error                 17.7112 %
Root relative squared error             44.3281 %
Total Number of Instances              698

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0.956     0.066     0.965      0.956    0.960      0.886    0.962     0.966     2
               0.934     0.044     0.918      0.934    0.926      0.886    0.962     0.923     4
Weighted Avg.  0.948     0.059     0.949      0.948    0.949      0.886    0.962     0.951

=== Confusion Matrix ===

   a   b   <-- classified as
 437  20 |   a = 2
  16 225 |   b = 4
```



طبقه بندی به روش Navies Bayesian :

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

Correctly Classified Instances         680               97.4212 %
Incorrectly Classified Instances        18                2.5788 %
Kappa statistic                          0.9436
Mean absolute error                      0.0275
Root mean squared error                  0.1586
Relative absolute error                  6.091  %
Root relative squared error             33.3648 %
Total Number of Instances              698

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.967    0.012    0.993      0.967    0.980      0.944    0.992     0.996     2
                 0.988    0.033    0.941      0.988    0.964      0.944    0.992     0.985     4
Weighted Avg.    0.974    0.019    0.975      0.974    0.974      0.944    0.992     0.992

=== Confusion Matrix ===

   a   b   <-- classified as
 442  15 |   a = 2
   3 238 |   b = 4
```

Accuracy=(442+238)/(442+238+15+3)=0.974

2-4- ارزیابی

مقایسه دوروش :

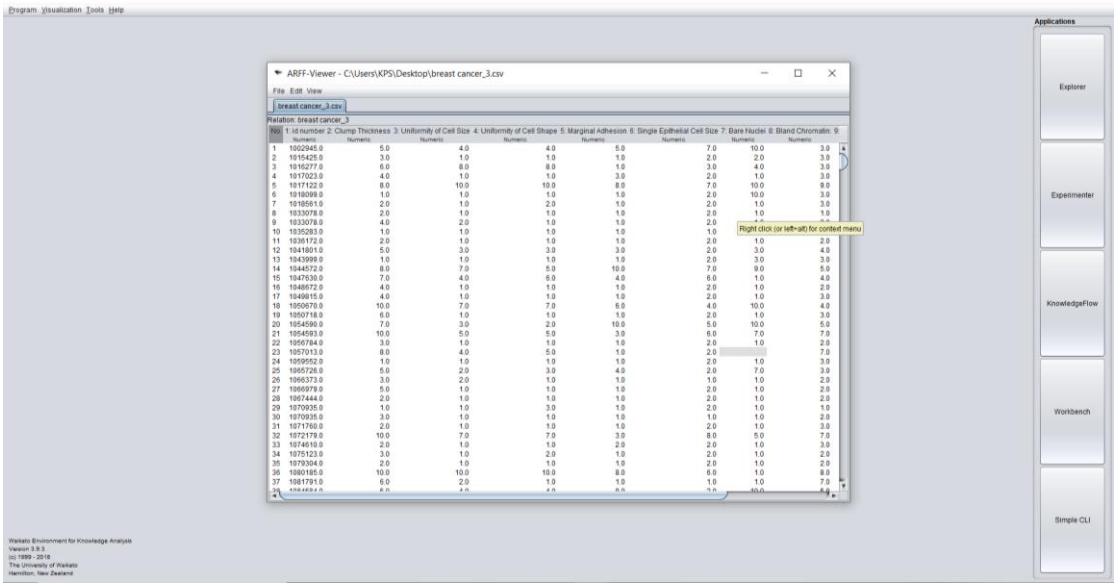
در کل برای این دیتاست روش دوم(Navies Bayesian) بهتر عمل کرده و با اعمال داده test توانسته است حدود97 درصد داده های train را درست بگوید این درحالیست که الگوریتم اول(j48) به 94درصد از داده های train پاسخ صحیح داده است .

	Accuracy	Precision	Recall	F-Measure
Navies Bayesian	0.974	0.993	0.967	0.980
J48	0.947	0.965	0.956	0.960

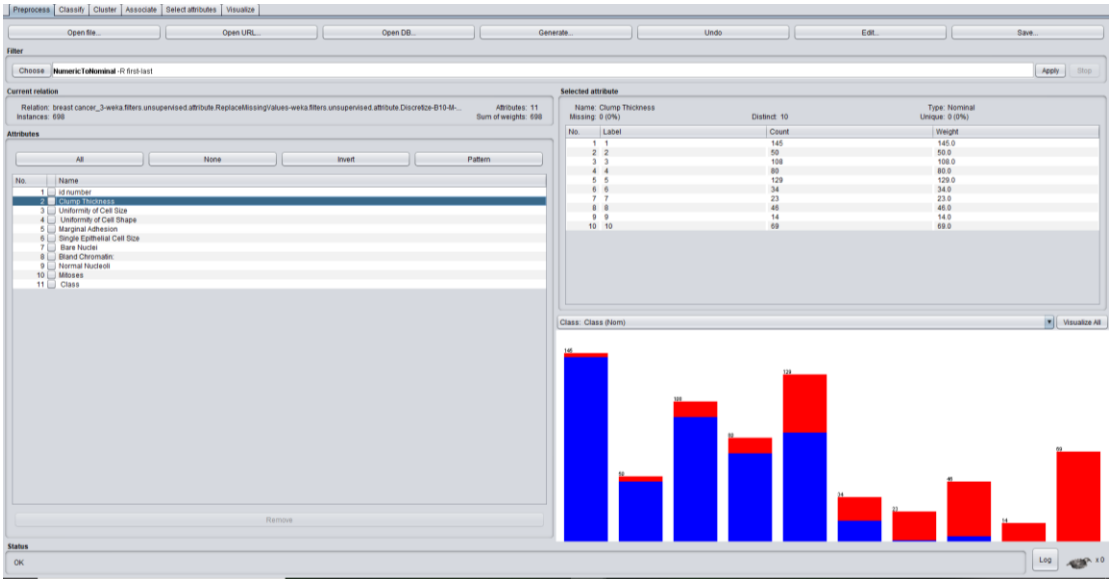
پیوست

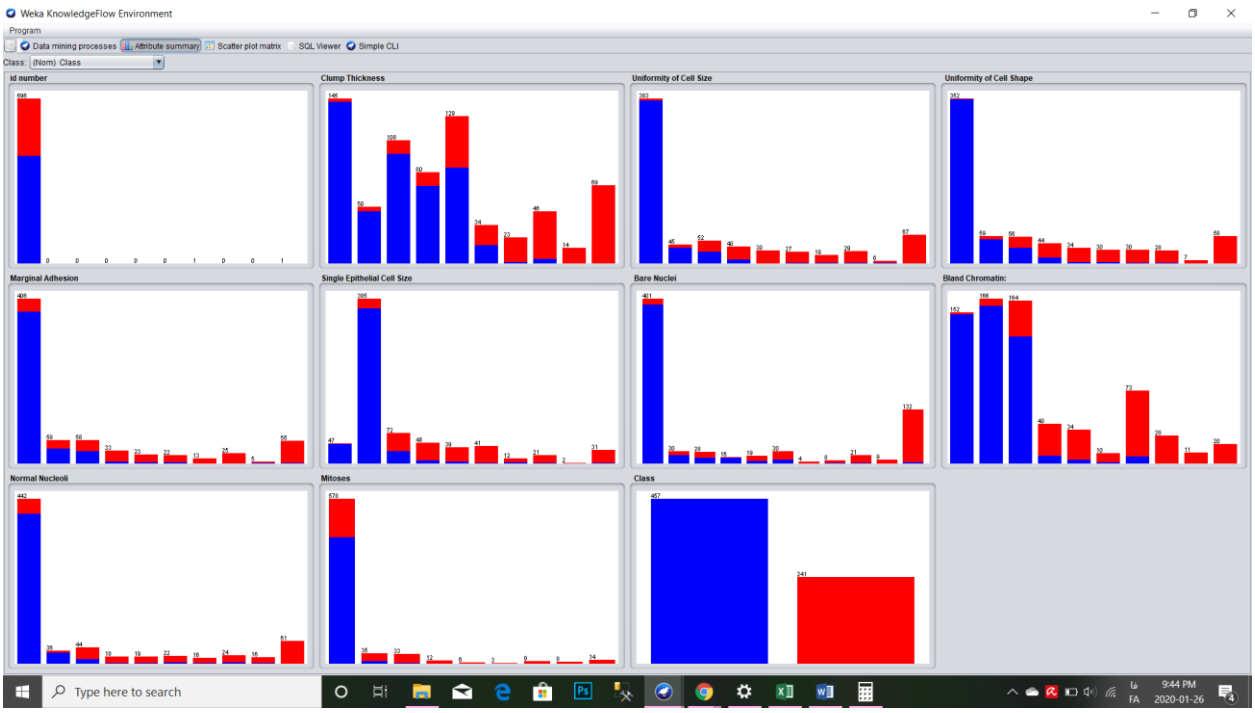
نحوه کار با وکا

در ابتدای کار فایل اصلی داده ای که قصد داشتیم با آن کار کنیم به فرمت data. ذخیره شده بود ، بنابراین ابتدا از داخل excel به فرمت csv. ذخیره کردیم و سپس در خود برنامه weka از طریق مسیر ArffViewer <- tools فایل را باز میکنیم و سپس به فرمت arff ذخیره میکنیم :



سپس مرحله پیش پردازش داده ها را در پیش داریم که من از فیلتر های numerictonuminal و normalize و descretize و نیز replacemissingvalues استفاده کردم تا داده هایم پاک سازی شوند و نیز به فرمت numinal در بیایند تا راحتتر قابل کلاس بندی باشند.


Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
NumericToNominal -R first-last
Apply
Stop
Current relation
Relation: breast cancer_3-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Discretize-B10-M-...
Instances: 698
Attributes: 11
Sum of weights: 698
Selected attribute
Name: Clump Thickness
Missing: 0 (0%)
Distinct: 10
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
Weight
1 1 145 145.0
2 2 50 50.0
3 3 108 108.0
4 4 80 80.0
5 5 129 129.0
6 6 34 34.0
7 7 23 23.0
8 8 46 46.0
9 9 14 14.0
10 10 69 69.0
Attributes
All
None
Invert
Pattern
No.
Name
1 id number
2 Clump Thickness
3 Uniformity of Cell Size
4 Uniformity of Cell Shape
5 Marginal Adhesion
6 Single Epithelial Cell Size
7 Bare Nuclei
8 Bland Chromatin
9 Normal Nucleoli
10 Mitoses
11 Class
Remove
Class: Class (Nom)
Visualize All
Status
OK
Log
x 0


Weka KnowledgeFlow Environment
Program
Data mining processes
Attribute summary
Scatter plot matrix
SQL Viewer
Simple CLI
Class: (Nom) Class
id number
Clump Thickness
Uniformity of Cell Size
Uniformity of Cell Shape
Marginal Adhesion
Single Epithelial Cell Size
Bare Nuclei
Bland Chromatin:
Normal Nucleoli
Mitoses
Class
Type here to search
9:44 PM
2020-01-26