

**دانشگاه صنعتی اصفهان**

**دانشکده برق و کامپیوتر**

**عنوان**

**تمرین شماره 1 درس یادگیری ماشین**

**استاد درس:**

**دکتر پالهنگ**

**دانشجو:**

**داریوش حسن پور**

*هوش مصنوعی (9308164)*

1393

# معرفی مجموعه داده

* **نام:** تأیید اعتبار (Credit Approval)
* **منبع:** quinlan@cs.su.oz.au
* **استفاده های قبلی:**
  + "Simplifying decision trees", Int J Man-Machine Studies 27, Dec 1987, pp. 221-234.
  + "C4.5: Programs for Machine Learning", Morgan Kaufmann, Oct 1992
* **توضیحات:**

فایل مربوط به مجموعه داده ای شامل اصلاعاتی راجع به کارت های اعتباری هست. که با توجه به توضیحاتی که در راجع به فایل آمده نام ویژگی برای حفاظت از داده های اصلی تغییر داده شده اند.

* **تعداد نمونه ها :** 690 عدد
* **تعداد ویژگی ها:** 15 عدد (16 تا به همراه ویژگی کلاس)
* **تعداد کلاس ها:** 2 عدد (+ یا -)

### ویژگی های و محدودیت دامنه ای آنها:

* + A1: { b, a }
  + A2: پیوسته
  + A3: پیوسته
  + A4: { u, y, l, t }
  + A5: { g, p, gg }
  + A6: { c, d, cc, i, j, k, m, r, q, w, x, e, aa, ff }
  + A7: { v, h, bb, j, n, z, dd, ff, o }
  + A8: پیوسته
  + A9: { t, f }
  + A10: { t, f }
  + A11: پیوسته
  + A12: { t, f }
  + A13: { g, p, s }
  + A14: پیوسته
  + A15: پیوسته
  + A16: { +, - } (ویژگی کلاس)

### توزیع کلاس ها:

* + + : 307 (44.5%)
  + - : 383 (55.5%)

### آماده سازی داده ها برای استفاده از داده در نرم افزار وکا:

برای اینکه داده های موجود در UCI را بتوانیم در نرم افزار وکا استفاده کنیم باید بر داده موجود در فایل crx.data گرفته شده از UCI اطلاعات زیر را اضافه کنیم و فایل جدید را با پسوند .arff ذخیره کنیم:

@RELATION credit

@ATTRIBUTE A1 {b,a}

@ATTRIBUTE A2 REAL

@ATTRIBUTE A3 REAL

@ATTRIBUTE A4 {u,y,l,t}

@ATTRIBUTE A5 {g,p,gg}

@ATTRIBUTE A6 {c,d,cc,i,j,k,m,r,q,w,x,e,aa,ff}

@ATTRIBUTE A7 {v,h,bb,j,n,z,dd,ff,o}

@ATTRIBUTE A8 REAL

@ATTRIBUTE A9 {t,f}

@ATTRIBUTE A10 {t,f}

@ATTRIBUTE A11 REAL

@ATTRIBUTE A12 {t,f}

@ATTRIBUTE A13 {g,p,s}

@ATTRIBUTE A14 REAL

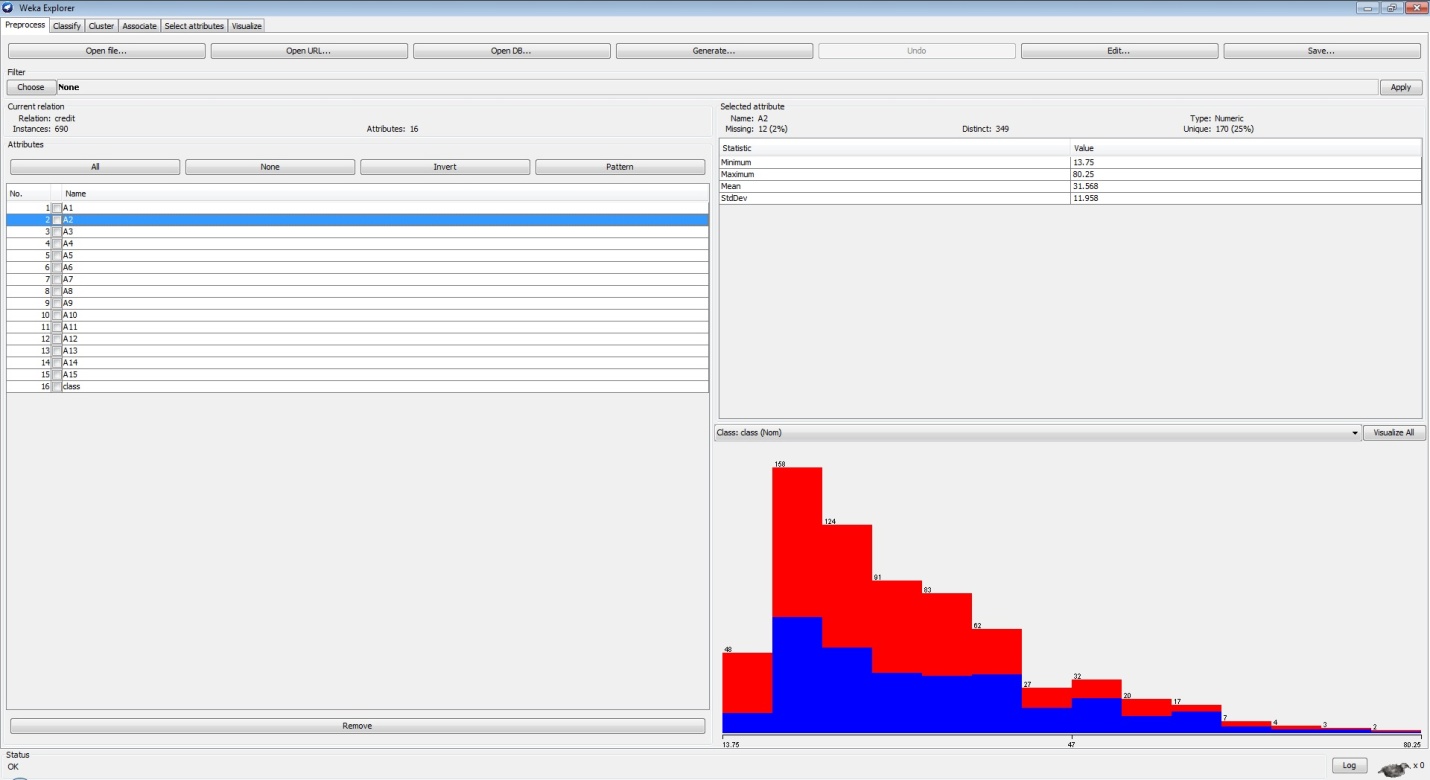
@ATTRIBUTE A15 REAL

@ATTRIBUTE class {+, -}

@DATA

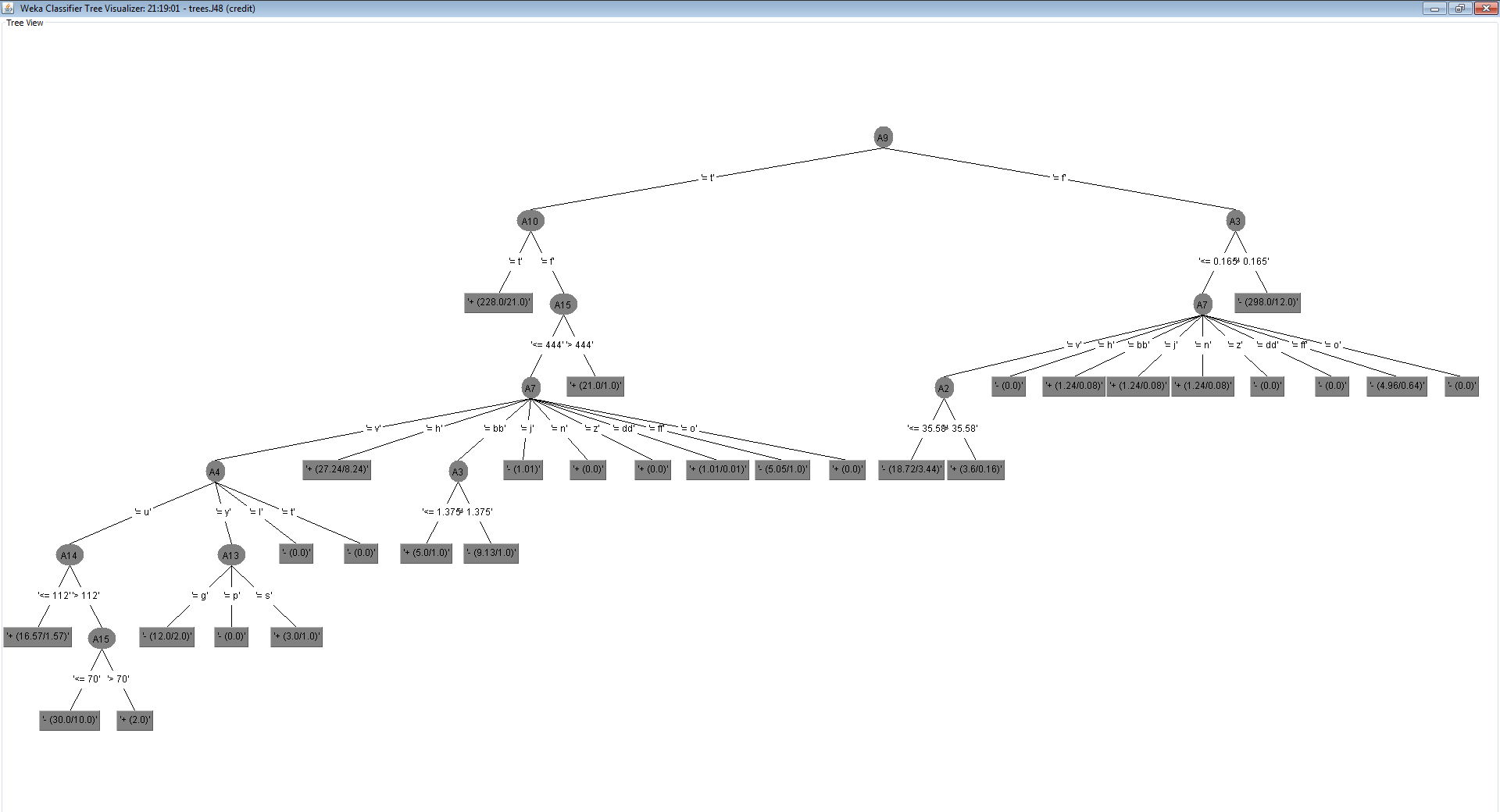
# توزیع داده ها در وکا

بعد از اینکه فایل .arff رو در وکا لود کردیم میتوانیم توزیع داده ای مربوط به هر ویژگی رو مشاهده کنیم. به عنوان مثال در ویژگی زیر به سادگی مشاهده میشود که هرکدام از مقادیر ویژگی دارای چه توزیعی است.

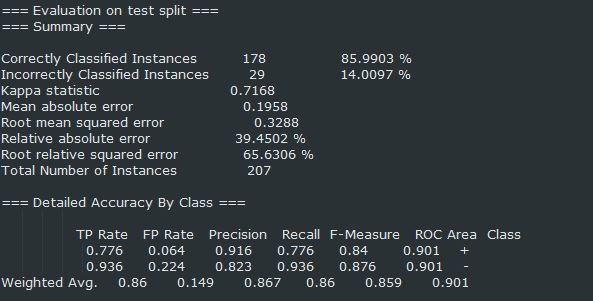


# ساختن درخت تصمیم با استفاده از الگوریتم C4.5 در وکا

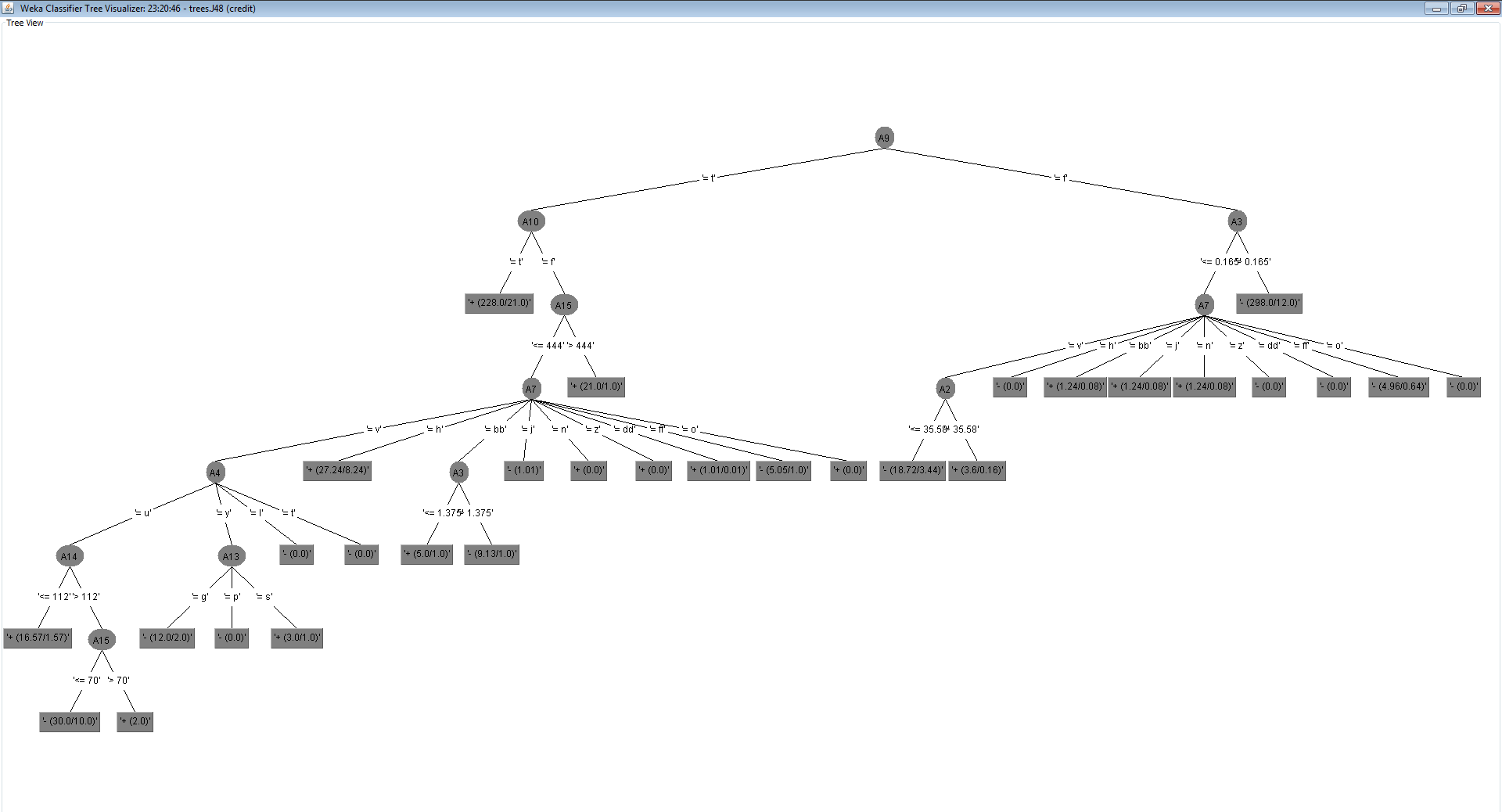
در صورتی که درخت با یک مجموعه آموزشی(70%) و یک مجموعه تست(30%) درخت تصمیم زیر تشکیل می‌شود.



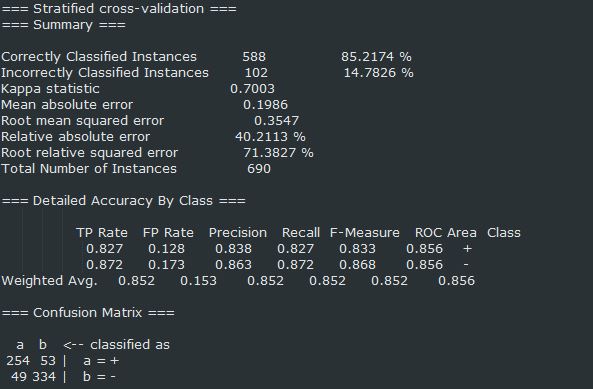
که نتایج آن در مجموعه تست به شرح زیر است:



و درخت با تصدیق 3 تایی به شرح زیر است:



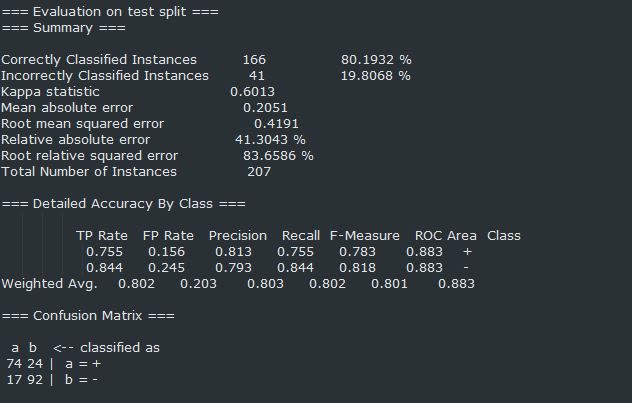
که نتایج آن در مجموعه تست به شرح زیر است:

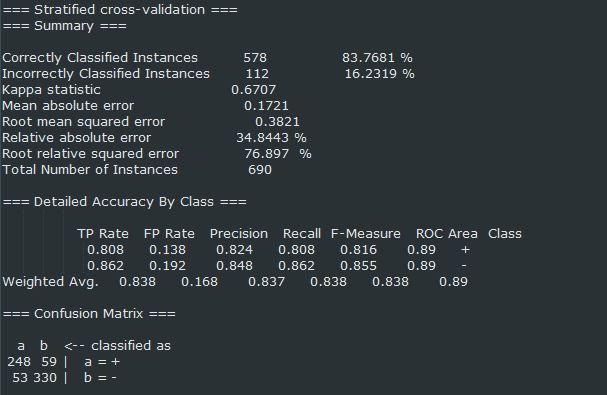


# آموزش با استفاده از شبکه عصبی

به تعداد 3 عدد نمونه ویژگی A2 آنها از مقدار اصلی تغییر پیدا کرده و به تعداد 12 عدد از نمونه ها مقدار مجهول ویژگی A2 آنها به طور تصادفی مقدار دهی شده اند؛ با تغییر ویژگیA2 به تعداد 15عدد از داده ها که شامل پر کردن تصادفی داده هایی که مقداری برای این ویژگی نداشتن تغییراتی را در داده ها ایجاد کردیم.

در صورتی که شبکه چند لایه با یک مجموعه آموزشی(70%) و یک مجموعه تست(30%) نتایج زیر را داریم:



و شبکه چند لایه با تصدیق 3 تایی به شرح زیر است:

# مقایسه و نتیجه

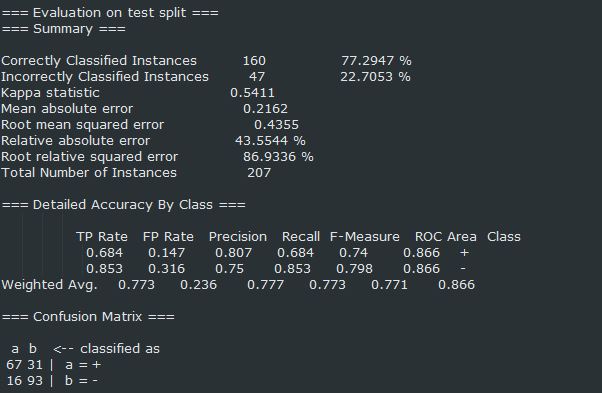
برای مقایسه یادگیری های انجام شده، فاکتور های ROC که وکا در خروجی آورده است را در نظر میگیریم که در جدول زیر آمده است.

|  |  |  |
| --- | --- | --- |
| شبکه عصبی | درخت تصمیم | الگوریتم  روش آموزش |
| 0.883 | 0.901 | **مجموعه آموزشی/تست** |
| 0.890 | 0.856 | **تصدیق مقاطع 3 تایی** |

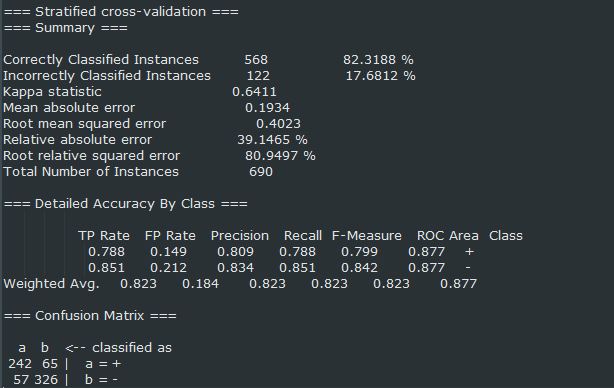
همانطور که میبینیم درخت تصمیم با روش آموزشی "مجموعه آموزشی/تست" مقدار ROC بالاتری نسبت به بقیه دارند بنابراین برای یادگیری این مفهوم هدف مناسب است. البته لازم به ذکر است بنابه اینکه در استفاده از شبکه عصبی داده های آموزشی دست کاری شده اند برای همین نمیتوان گفت که شبکه عصبی برای یادگیری این مفهوم مناسب نیست چرا که داده ها دارای نویز بوده اند؛ بنابراین در یک تلاش دیگر با **داده های اصلی بدون دست کاری شده** سعی شده که شبکه عصبی را آموزش داده که در بخش بعدی به آورده شده و مقایسه ای با نتایج درخت تصمیم ساخته شده در قسمت قبل انجام شده است.

# شبکه عصبی با داده های بدون دست کاری شده

در صورتی که شبکه چند لایه با یک مجموعه آموزشی(70%) و یک مجموعه تست(30%) نتایج زیر را داریم:



و شبکه چند لایه با تصدیق 3 تایی به شرح زیر است:



# مقایسه و نتیجه کلی

برای مقایسه یادگیری های انجام شده، فاکتور های ROC که وکا در خروجی آورده است را در نظر میگیریم که در جدول زیر آمده است.

|  |  |  |  |
| --- | --- | --- | --- |
| شبکه عصبی(بدون دستکاری داده ای) | شبکه عصبی(با دستکاری داده ای) | درخت تصمیم | الگوریتم  روش آموزش |
| 0.866 | 0.883 | 0.901 | **مجموعه آموزشی/تست** |
| 0.877 | 0.890 | 0.856 | **تصدیق مقاطع 3 تایی** |

باوجود اینکه دادهای بدون دست کاری شده توسط شبکه عصبی آموزش دیده شدند ولی مقدار ROC آن در هر دو روش آموزشی کمتر از زمانی است که داده های شبکه عصبی دست کاری شده اند؛ همانطور که پیشتر گفته شد به تعداد 3 عدد نمونه ویژگی A2 آنها از مقدار اصلی تغییر پیدا کرده و به تعداد 12 عدد از نمونه ها مقدار مجهول ویژگی A2 آنها به طور تصادفی مقدار دهی شده اند؛ که همانطور که نتایج فوق نشان میدهند تغییرات اعمال شده باعث شده اند شبکه عصبی مفهوم را بهتر از زمانی که داده بدون دست کاری استفاده شده اند را یادگرفته است. ولی باز نیز همانطور که میبینیم درخت تصمیم با روش آموزشی "مجموعه آموزشی/تست" مقدار ROC بالاتری نسبت به بقیه دارند بنابراین برای یادگیری این مفهوم هدف مناسب است.