به نام خدا

حل مسئله‌ی طبقه‌بندی به وسیله درخت تصمیم

"درس مبانی علم داده‌ها"

استاد مربوطه: دکتر موسی گلعلی‌زاده

پژوهشگر: ساجده لشگری

آبان ۹۸

داده‌های مورد استفاده، مربوط به قسمت منابع انسانی یکی از شرکت‌های بزرگِ آمریکایی است.

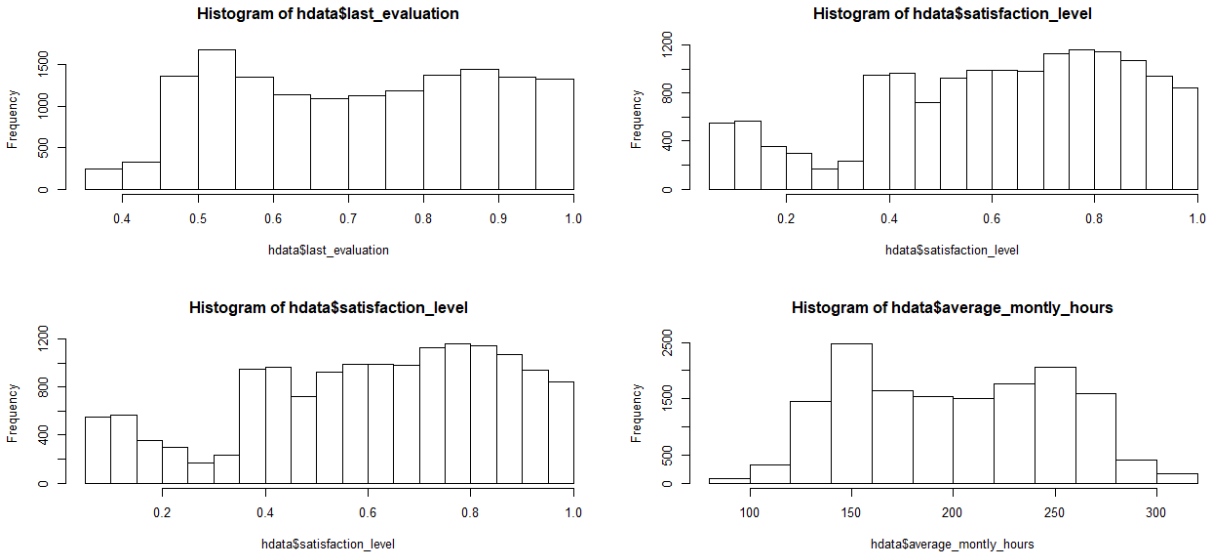
تعداد کارمندان این شرکت ۱۴۹۹۹ نفر می‌باشد که در بین آن‌ها ۸۰% به عنوان نمونه‌ی آموزش و ۲۰% نمونه آزمایش در نظر گرفته شده‌اند.

همچنین صفت‌های مربوط به این نمونه‌ها، سطح رضایت‌مندی (متغیر عددی بین ۰ تا ۱)، میانگین ساعات کاری در ماه (متغیر عددی بین ۹۶ تا ۳۱۰)، ترفیع شغلی (متغیر رسته‌ای با دو مقدار ۰و۱ که به ترتیب به معنی ارتقا گرفتن یا نگرفتن است)، تعداد پروژه‌های انجام داده‌شده توسط کارمند (متغیر عددی بین ۲ تا ۷)، رخدادن حادثه در محل کار برای هر کارمند (متغیر رسته‌ای ۰ و ۱)، امتیازات به‌دست آمده از آخرین ارزیابی توسط کارفرما (متغیر عددی بین ۰ تا ۱)، سال‌های گذرانده در آن شرکت (متغیر عددی بین ۲ تا ۱۰)، بخشی که هر کارمند در آن کار می‌کند (متغیر رسته‌ای ۹سطحی)، درآمد (متغیر رسته‌ای ۳ سطحی/ کم، متوسط، زیاد) و متغیر آخر (که به عنوان متغیر پاسخ در نظر گرفته شده و هدف پیش‌بینی آن است) ماندن یا استفاء از شغل(متغیر رسته‌ای دو سطحی) می‌باشند.

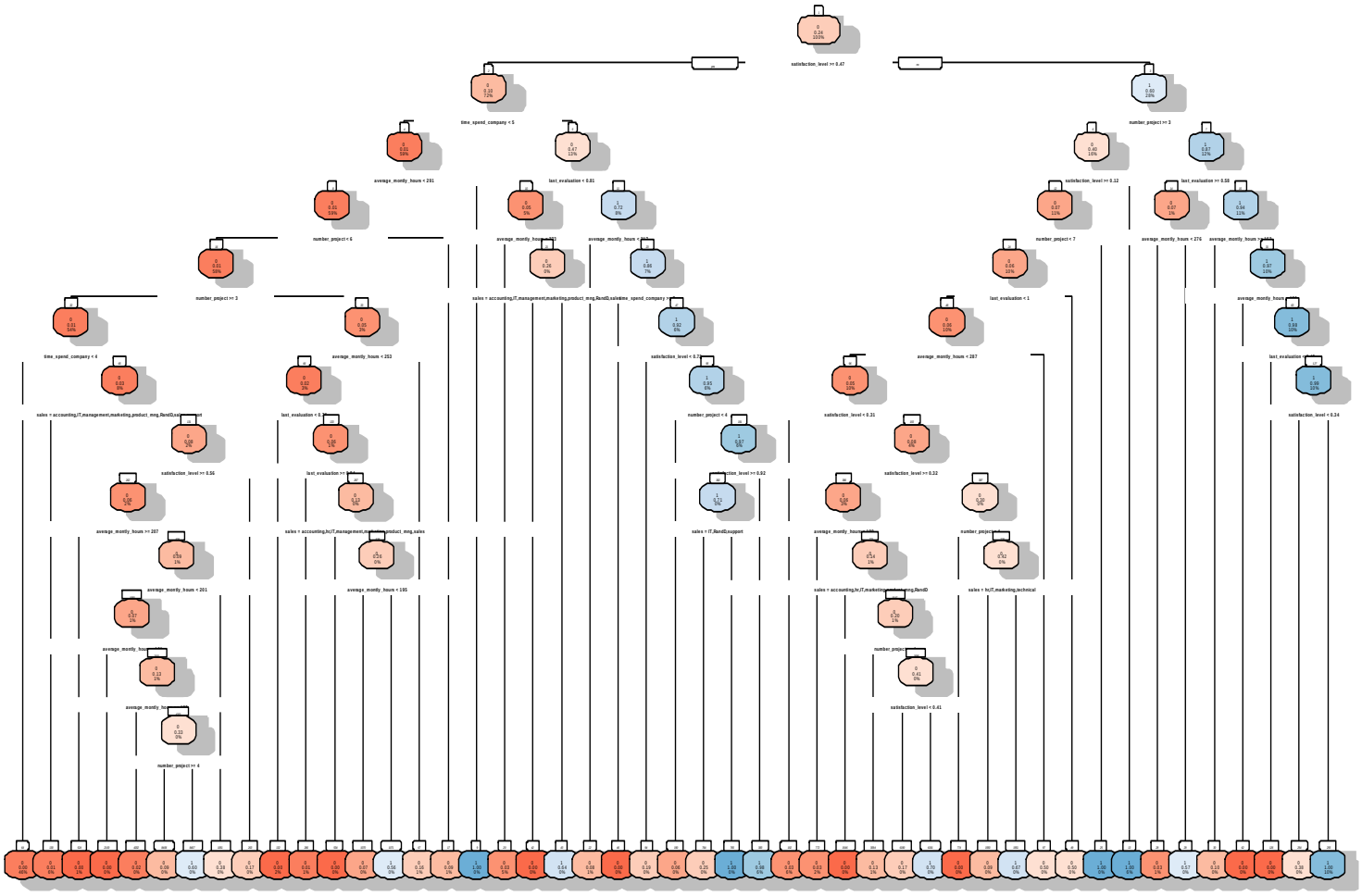
خلاصه‌ای از داده‌ها به صورت زیر است:

```
satisfaction_level last_evaluation  number_project  average_montly_hours time_spend_company
Min.   :0.0900     Min.   :0.3600   Min.   :2.000   Min.   : 96.0        Min.   : 2.000
1st Qu.:0.4400     1st Qu.:0.5600   1st Qu.:3.000   1st Qu.:156.0        1st Qu.: 3.000
Median :0.6400     Median :0.7200   Median :4.000   Median :200.0        Median : 3.000
Mean   :0.6128     Mean   :0.7161   Mean   :3.803   Mean   :201.1        Mean   : 3.498
3rd Qu.:0.8200     3rd Qu.:0.8700   3rd Qu.:5.000   3rd Qu.:245.0        3rd Qu.: 4.000
Max.   :1.0000     Max.   :1.0000   Max.   :7.000   Max.   :310.0        Max.   :10.000

Work_accident       left        promotion_last_5years        sales        salary
Min.   :0.0000   Min.   :0.0000   Min.   :0.00000      sales      :4140   high  :1237
1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000      technical  :2720   low   :7316
Median :0.0000   Median :0.0000   Median :0.00000      support    :2229   medium:6446
Mean   :0.1446   Mean   :0.2381   Mean   :0.02127      IT         :1227
3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000      product_mng: 902
Max.   :1.0000   Max.   :1.0000   Max.   :1.00000      marketing  : 858
                                                       (Other)    :2923
```

باتوجه به خلاصه‌ی بالا، عدم وجود داده گمشده و دور افتاده مشاهده می‌شود. همچنین نمودارهای هیستوگرام داده‌های عددی در زیر نمایش داده شده‌اند که باتوجه به نمودارها، می‌توان گفت توزیع آن‌ها نرمال نیست.



در مرحله بعد مدل‌سازی با استفاده از درخت تصمیم انجام شده و سپس برای متغیر پاسخ (ماندن یا ترک کردن شغل (استعفاء از شغل)) پیش بینی انجام شده است.

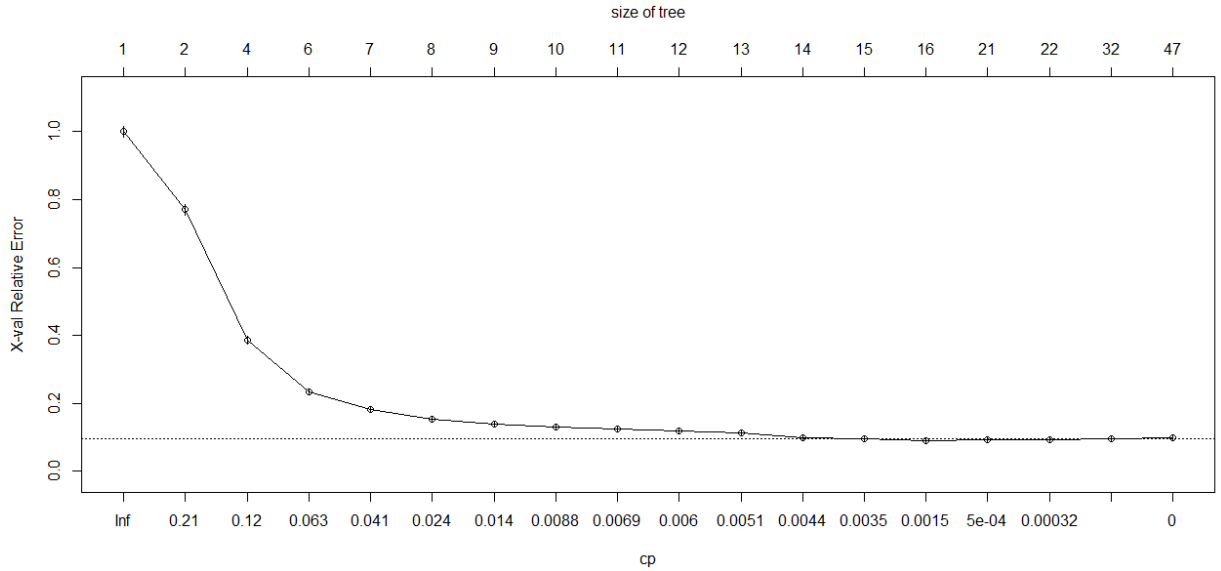


نمودار ۱

در ساختن مدل تنها از ۶ صفت در بین ۹ صفت نام برده شده در بالا، استفاده شده (در واقع انتخاب ویژگی (feature selection) انجام شده است) که می‌توان این نتایج را در جدول زیر مشاهده کرد.

```
Classification tree:
rpart(formula = left ~ ., data = train, method = "class", control = rpart.control(cp = 0))

Variables actually used in tree construction:
[1] average_montly_hours last_evaluation      number_project       sales
[5] satisfaction_level   time_spend_company

Root node error: 2816/11971 = 0.23524

n= 11971

           CP nsplit rel error   xerror      xstd
1  2.2976e-01      0  1.000000 1.000000 0.0164796
2  1.9229e-01      1  0.770241 0.770241 0.0149654
3  7.6882e-02      3  0.385653 0.385653 0.0111592
4  5.2202e-02      5  0.231889 0.233310 0.0088490
5  3.2315e-02      6  0.179688 0.182173 0.0078689
6  1.7401e-02      7  0.147372 0.153764 0.0072546
7  1.1009e-02      8  0.129972 0.139205 0.0069148
8  7.1023e-03      9  0.118963 0.129261 0.0066713
9  6.7472e-03     10  0.111861 0.125000 0.0065638
10 5.3267e-03     11  0.105114 0.118253 0.0063894
11 4.9716e-03     12  0.099787 0.112926 0.0062479
12 3.9062e-03     13  0.094815 0.099432 0.0058723
13 3.1960e-03     14  0.090909 0.095526 0.0057585
14 7.1023e-04     15  0.087713 0.089844 0.0055884
15 3.5511e-04     20  0.084162 0.093395 0.0056954
16 2.9593e-04     21  0.083807 0.094105 0.0057165
17 7.1023e-05     31  0.080611 0.095526 0.0057585
18 0.0000e+00     46  0.079545 0.100142 0.0058927
```

جدول۱



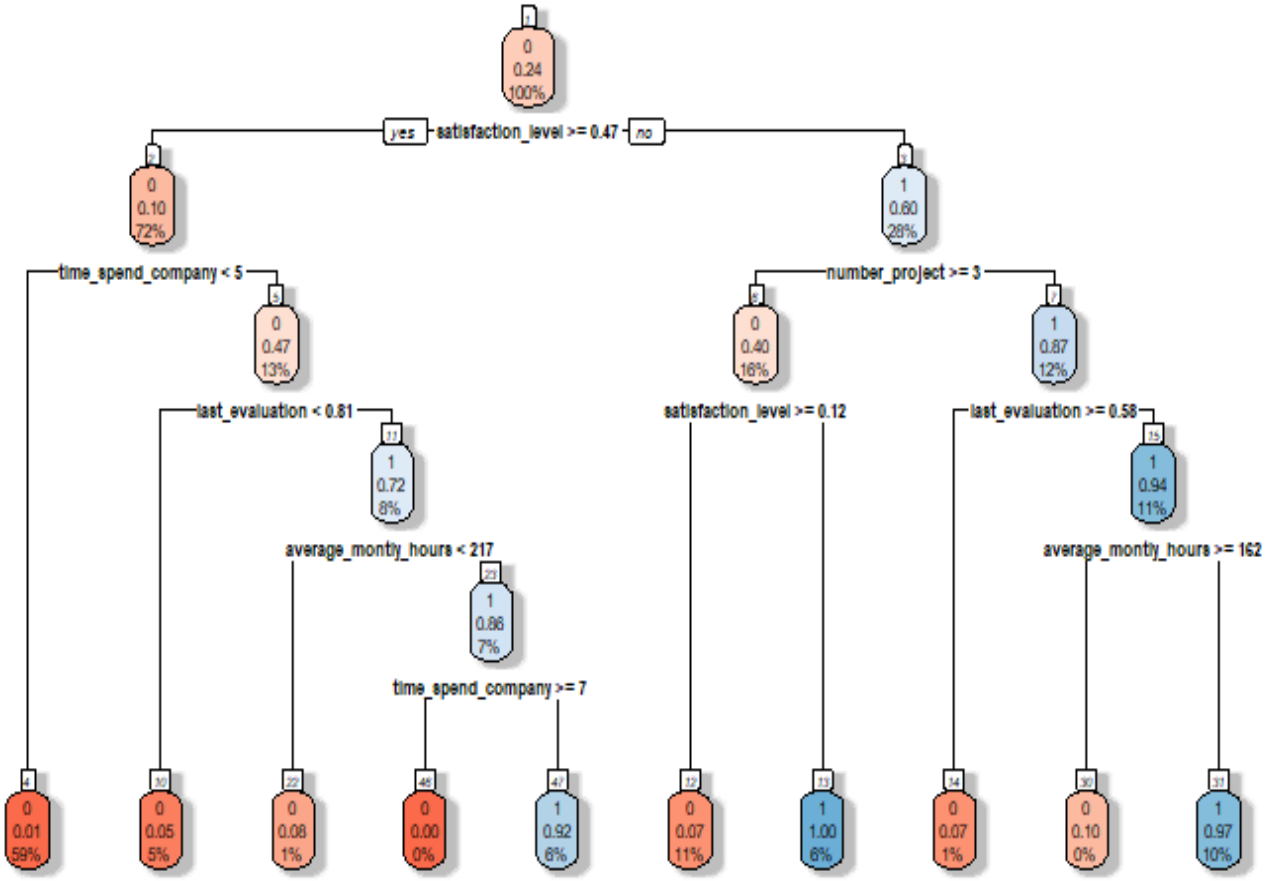
نمودار۲

همانطور که در نمودار اول مشاهده می‌شود، تعداد splitها و شاخه‌ها خیلی زیاد می‌باشند و به نظر می‌رسد بیش برازش صورت گرفته است. در نتیجه درخت، نیاز به هرس شدن دارد.

همچنین تفسیر از روی این درخت کار دشواری است. در مراحل بعدی، پس از انجام هرس کردن به تفسیر نتایج پرداخته می‌شود.

در اینجا، درخت به ۲ روش pre-pruning و post-pruning هرس شده است. ابتدا با روش pre-pruning هرس کردن صورت گرفته و قیدهای در نظر گرفته شده (که تعیین آن‌ها با خودمان بوده) به صورت زیر است:

minsplit = 100 cp = 0 ,maxdepth = 8,



سپس نتایج زیر حاصل شده است.

```
Classification tree:
rpart(formula = left ~ ., data = train, method = "class", control = rpart.control(cp = 0,
    maxdepth = 8, minsplit = 100))

Variables actually used in tree construction:
[1] average_montly_hours last_evaluation      number_project       satisfaction_level
[5] time_spend_company

Root node error: 2816/11971 = 0.23524

n= 11971

        CP nsplit rel error  xerror      xstd
1 0.229759      0   1.00000 1.00000 0.0164796
2 0.192294      1   0.77024 0.77024 0.0149654
3 0.076882      3   0.38565 0.38565 0.0111592
4 0.052202      5   0.23189 0.23331 0.0088490
5 0.032315      6   0.17969 0.18288 0.0078836
6 0.017401      7   0.14737 0.15447 0.0072707
7 0.011009      8   0.12997 0.13814 0.0068892
8 0.000000      9   0.11896 0.12642 0.0065999
```

همانطور که مشاهده می‌شود، تعداد صفت‌های استفاده شده و تعداد splitها در این مدل کمتر از مدل قبل است.

ابتدا به تفسیر نتایج پرداخته و در آخر این مدل‌ها از حیث دقت باهم مقایسه می‌شوند.

چند نمونه از نتایج به‌دست آمده به صورت زیر می‌باشد:

- ۵۹% از کارمندانی که کمتر از ۵ سال در شرکت کار کرده‌اند و سطح رضایت‌مندی آن‌ها بیشتر از ۴۷% بوده است، از کار خود استعفا داده‌اند. (مربوط به اولین برگ از سمت چپ)

- ۵% از کارمندانی که آخرین نمره‌ی ارزیابی آن‌ها کمتر از ۸۱% بوده و بیشتر از ۵ سال در شرکت کار کرده‌اند و سطح رضایت‌مندی آن‌ها بیشتر از ۴۷% بوده است، از کار خود استعفا داده‌اند. (مربوط به دومین برگ از سمت چپ)

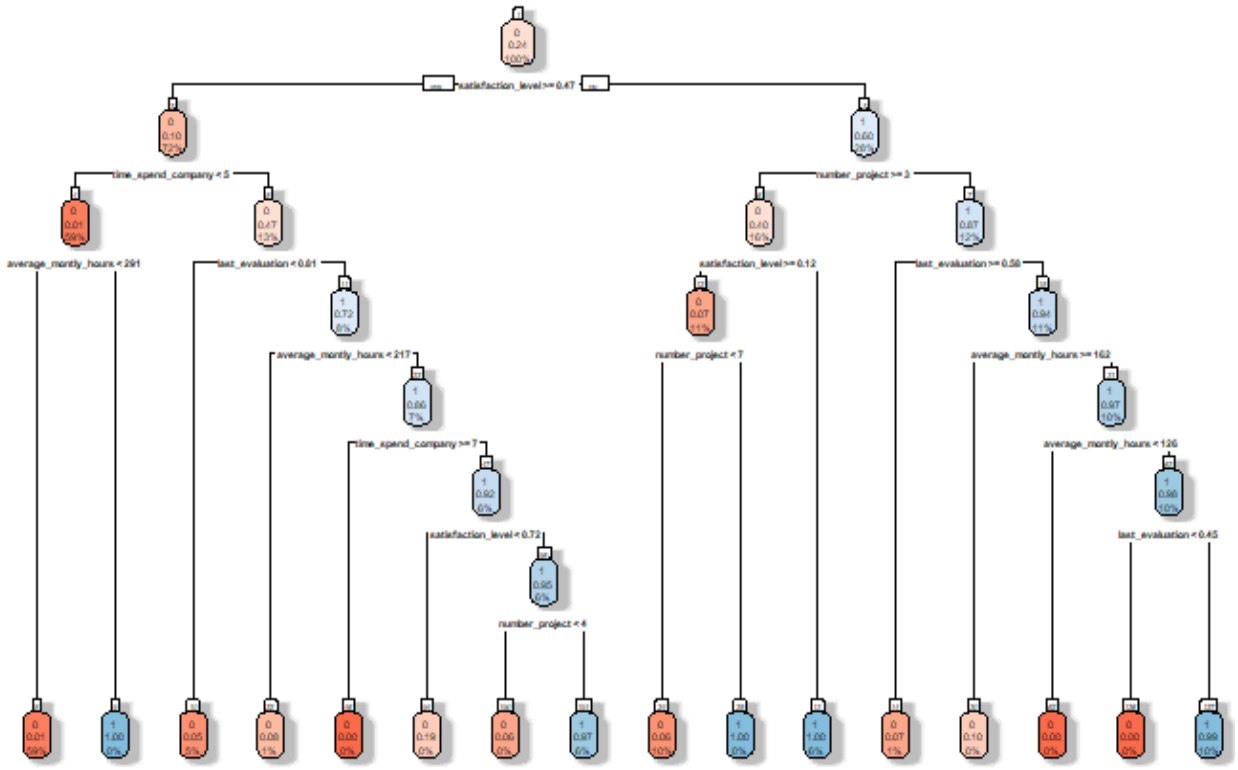
- ۶% از کارمندانی که آخرین نمره‌ی ارزیابی آن‌ها بیشتر از ۸۱% بوده و به طور میانگین بیشتر از۲۱۷ ساعت در ماه و به طور کلی بین ۵ تا ۷ سال در شرکت کار کرده‌اند همچنین سطح رضایت‌مندی آن‌ها بیشتر از ۴۷% بوده است ، به کار کردن در آن شرکت ادامه داده‌اند. (مربوط به پنجمین برگ از سمت چپ)

- ۱۰% از کارمندانی که به طور میانگین کمتر از ۱۶۲ ساعت در ماه کار کرده‌اند و آخرین نمره‌ی ارزیابی آن‌ها کمتر از ۵۸% بوده، همچنین کمتر از ۳ پروژه توسط آن‌ها انجام شده و سطح رضایت‌مندی آن‌ها کمتر از ۴۷% بوده است، به کار کردن در آن شرکت ادامه داده‌اند. (مربوط به اولین برگ از سمت راست)

در مرحله بعد با استفاده از نتایج به‌دست آمده در جدول۱ و نمودار۲ تعداد splitهای مناسب تعیین می‌شود و post-pruning به وسیله آن انجام می‌شود.

با توجه به کوچکترین مقدار xerror در جدول یادشده و نقطه‌ی شکستگی در نمودار۲ (و برخورد آن با خطِ نقطه چین) می‌توان گفت split مناسب در نقطه ۱۵ قرار دارد پس به splitهای بعد از آن نیازی نیست و وجود آن‌ها باعث بیش برازش می‌شود.

درخت حاصل از این مدل به‌صورت زیر می‌باشد:



همچنین نتایج و تفاسیر حاصل از آن در زیر مشاهده می‌شود.

```
Classification tree:
rpart(formula = left ~ ., data = train, method = "class", control = rpart.control(cp = 0))

Variables actually used in tree construction:
[1] average_montly_hours last_evaluation      number_project       satisfaction_level
[5] time_spend_company

Root node error: 2816/11971 = 0.23524

n= 11971

           CP nsplit rel error   xerror      xstd
1  0.22975852      0  1.000000 1.000000 0.0164796
2  0.19229403      1  0.770241 0.770241 0.0149654
3  0.07688210      3  0.385653 0.385653 0.0111592
4  0.05220170      5  0.231889 0.233310 0.0088490
5  0.03231534      6  0.179688 0.182173 0.0078689
6  0.01740057      7  0.147372 0.153764 0.0072546
7  0.01100852      8  0.129972 0.139205 0.0069148
8  0.00710227      9  0.118963 0.129261 0.0066713
9  0.00674716     10  0.111861 0.125000 0.0065638
10 0.00532670     11  0.105114 0.118253 0.0063894
11 0.00497159     12  0.099787 0.112926 0.0062479
12 0.00390625     13  0.094815 0.099432 0.0058723
13 0.00319602     14  0.090909 0.095526 0.0057585
14 0.00071023     15  0.087713 0.089844 0.0055884
```

دو نمونه از تفسیرهای به‌دست آمده با استفاده از مدل آخر، به‌صورت زیر می‌باشد:

- ۱۰% از کارمندانی که به طور میانگین بین ۱۲۶ تا ۱۶۲ ساعت در ماه کار کرده‌اند و آخرین نمره‌ی ارزیابی آن‌ها بین ۴۵% تا ۵۶% بوده، همچنین کمتر از ۳ پروژه توسط آن‌ها انجام شده و سطح رضایت‌مندی آن‌ها کمتر از ۴۷% بوده است، به کار کردن در آن شرکت ادامه داده‌اند. (مربوط به اولین برگ از سمت راست)

- ۵۹% از کارمندانی که کمتر از ۵ سال و به‌طور میانگین در ماه کمتر از ۲۹۱ ساعت در شرکت کار کرده‌اند، همچنین سطح رضایت‌مندی آن‌ها بیشتر از ۴۷% بوده است، از کار خود استعفا داده‌اند. (مربوط به اولین برگ از سمت چپ)

دقت‌های مربوط به ۳ مدل، به شرح زیر می‌باشد.

```
  base_accuracy accuracy_preprun accuracy_postprun
1     0.9689564        0.9666446         0.9719287
```

همانطور که مشاهده می‌شود دقت مدل آخر بیشتر از مدل‌های دیگر است. همچنین در مدل دوم با توجه به استفاده از قیدهای تعیین شده، دقت مدل نسبت به مدل هرس نشده کمتر شده است و می‌توان نتیجه گرفت قیدها خوب انتخاب نشده‌اند.

همچنین ماتریس در هم آمیختگی به صورت زیر می‌باشد:

مقادیر واقعی / مقادیر پیش بینی شده	۰	۱
۰	۲۲۶۶	۷۸
۱	۷	۶۷۷

با توجه به ماتریس بالا، ۲۲۶۶ نفر از کارمندان شرکت از شغل خود استفاء داده بودند و با استفاده از مدل نیز این افراد، استفاء داده شده پیش بینی شده‌اند. همچنین ۶۷۷ نفر از آن‌ها کارمندانی بوده‌اند که به شغل خود ادامه داده‌اند و توسط مدل نیز، به درستی پیش‌بینی شده‌اند.

به علاوه، ۷۸ نفر از کارمندان به شغل خود ادامه داده‌اند ولی باتوجه به مدل استفاء داده‌شده پیش‌بینی شده‌اند و ۷ نفر برعکس، استفاء داده‌اند ولی باتوجه به مدل به عنوان کسانی پیش‌بینی شده‌اند که به شغل خود ادامه داده‌اند.

در جدول زیر مقادیر دقت، صحت، فراخوانی و خطا قابل مشاهده است که با توجه به مقدار نزدیک به یکِ ۳ مورد اول و مقدار نزدیک به صفرِ خطا، همچنین نزدیک بودن خطای آزمایش و آموزش بهم دیگر، می‌توان نتیجه گرفت مدل انتخاب شده، عملکرد خوبی دارد.

همچنین قابل ذکر است این مقادیر مربوط به مدل آخر (یعنی post-pruning) می‌باشد.

accuracy for test	accuracy for train	error for test	error for train	precission for test	recall for test
0.972	0.979	0.028	0.020	0.989	0.896

نمودار ROC در زیر نمایش داده شده است. محور افقی نرخ FP (به اشتباه استفاء داده، پیش‌بینی شده‌اند) و محور عمودی نرخ TP (به درستی استفاء داده، پیش‌بینی شده‌اند) را نشان می‌دهد.

همانطور که مشاهده می‌شود، در ابتدای نمودار FPR تقریبا ثابت و نزدیک به صفر بوده و درحالی‌که TPR در حال افرایش و نزدیک شدن به یک است و بعد از آن مقدار TPR تقریبا روی ۰/۹ ثابت و مقدار FPR افزایش می‌یابد، از این‌رو می‌توان گفت بهترین برازش در نقطه سمت چپ نمودار که نزدیک به یک است، قرار دارد و باتوجه به مقدار AUC=0.95 (که مساحت زیر منحنی نمودار ROC است) مدلِ برازش‌شده، مدل خوبی ارزیابی می‌شود.