بسم الله الرحمن الرحیم

بکارگیری درخت تصمیم و شبکه عصبی در نرم افزار tableau

استاد: جناب دکتراکبر فرهودی نژاد

پژوهشگر:مسعود محمدی[۱]

اسم درس:داده کاوی و انبار دادها

دانشگاه پیام نور ری

(سند دوم)

[۱]*masood_mohammadics @yahoo.com*

چکیده- نوشته جاری؛ شامل بخش های معرفی، ابزارها و نحوه کاربرد آنها، عمل و تئوری ساخت درخت تصمیم با استفاده از rpart و توضیح شبکه عصبی ، درخت تصمیم و شبکه عصبی می باشد. در بخش معرفی یک دید کلی از کاری که می خواهیم انجام دهیم را توضیح داده است. در بخش ابزارها و نحوه کاربرد آنها، هر آنچه که برای انجام اعمال داده کاوی لازم است نام برده شده و نحوه نصب نرم افزارهای tableau و R نیز مرحله به مرحله توضیح داده شده است. در بخش عمل و تئوری ساخت درخت تصمیم با استفاده از rpart و توضیح شبکه عصبی به صورت کامل تمام دستورات مورد نیاز برای انجام تمرینات عملی در tableau توضیح داده شده است . در بخش درخت تصمیم دو نوع درخت تصمیم شامل درخت طبقه بند و درخت رگرسیون به صورت کامل توضیح داده شده و نحوه پیاده سازی آنها در نرم افزار tableau مرحله به مرحله توضیح داده شده است. سرانجام در بخش شبکه عصبی نحوه کاربرد random Forest در نرم افزار tableau توضیح داده شده است.

I. معرفی

در یک نگاه داده کاوی شامل مراحل زیر می باشد:

۱- شناسایی هدف: در این مرحله مشخص می شود کاربر چه چیزی را می خواهد و تا چه سطحی از اطلاعات را در نظر دارد که از پایگاه داده، اخذ نماید.

۲- انتخاب داده ها: در این مرحله باید داده ها بر مبنای معیارهای مشخص انتخاب گردند.

۳- آماده سازی داده ها: شکل قابل استفاده داه و شناسایی متغیرهای زائد وظیفه این مرحله از فرایند خواهد بود.

۴- ارزیابی داده ها: چهاچوب کلی این مرحله، معیارهای از قبیل نوع توزیع داده ها، ویژگی ها و ساختار پایگاه داده و شرایط کلی داده ها و ... می باشد.

۵- قالب بندی پاسخ: خروجی این بخش ، ارایه فرمت به شکل تصویر، نمودار، شبکه عصبی و... است.

۶- انتخاب ابزار: در این مرحله ابزارهای مناسب برای داده کاوی انتخاب می گردد.

۷- الگوسازی: فرایند داده کاوی به صورت اصلی از این مرحله آغاز می گردد که شامل جستجوی الگوها در مجموعه داده، طبقه بندی و ارزشیابی داده ها و .... می باشند.

۸- اعتبار سازی یافته ها: این مرحله، شامل آزمون کردن الگوها است.

۹- ارایه نتایج: نتایج این بخش، گزارش نهایی برای کاربر است.

۱۰- استفاده از نتایج: هدف اصلی داده کاوی استفاده از نتایج کشف شده برای تصمیم گیری، سیاست گذاری و پیش بینی به منظور ایجاد یک موقعیت بهتر و جدید می باشد.(نجات و اکبری،۱۳۸۷)

در این پژوهش از ابزار داده کاوی tableau و برای نمایش ترسیمات نیز از R package استفاده است. در بخش III این نرم افزارها توضیح داده شده اند.

بلوک بندی بازگشتی[۱] یکی از ابزارهای اساسی و پایه در داده کاوی[۲] می باشد. این ابزارها برای کاوش ساختار مجموعه داده، به منظور توسعه راحت قواعد تصمیمگیری برای پیش بینی نتایج قطعی(درخت کلاس بندی[۳]) یا پیوسته (درخت رگرسیون[۴])به ما کمک می کنند. در بخش III مدل سازی CART شرح داده شده است.

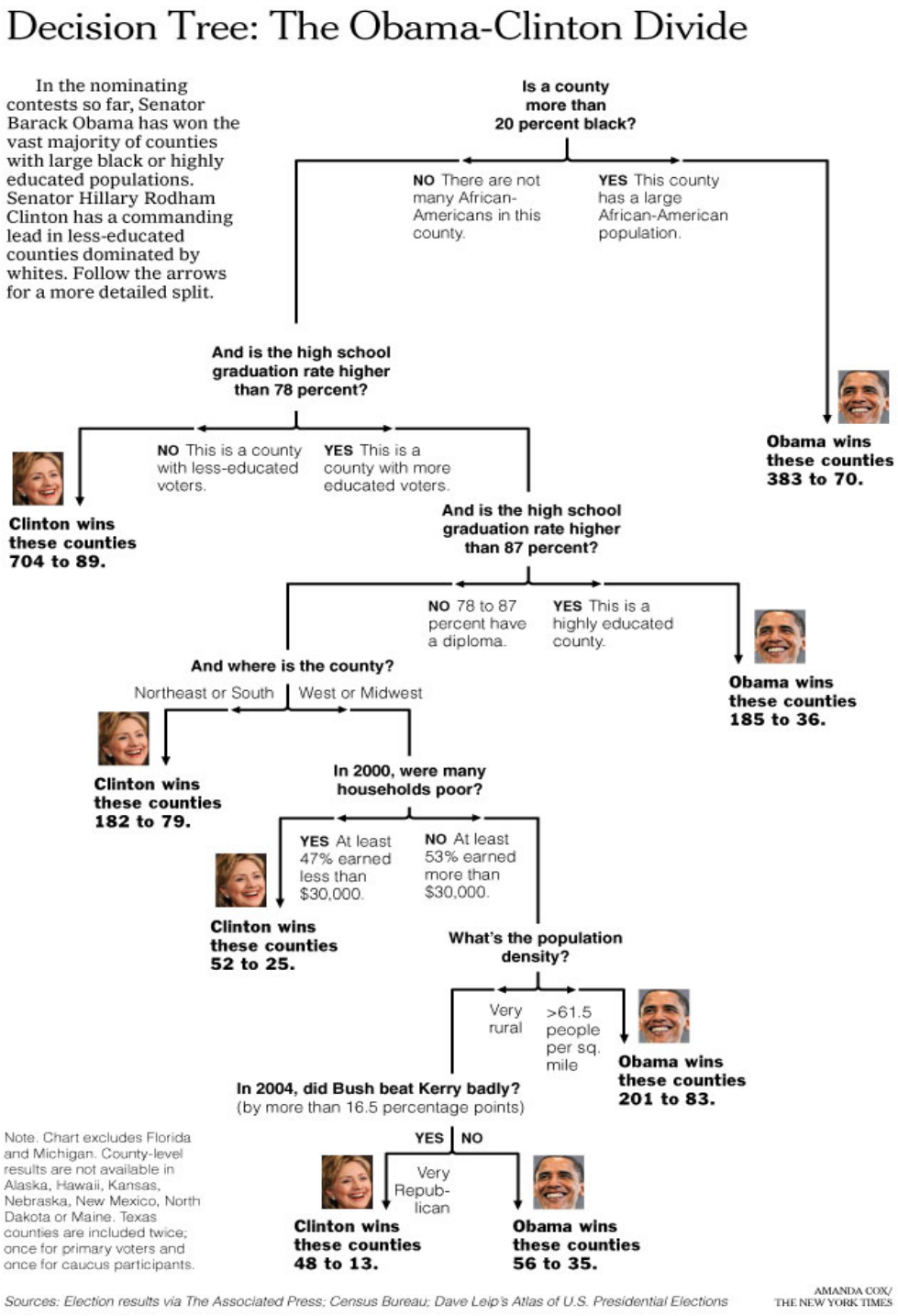
در مسائل یادگیری با دو موضوع سروکار داریم که عبارتند از: ۱- نحوه نمایش فرضیه ها و ۲- روشی که برای یادگیری بر می گزینیم. برای نمایش فرضیه ها از درخت تصمیم استفاده می شود. درخت تصمیم درختی است که در آن نمونه ها به نحوی دسته بندی می کنند که از ریشه به سمت پایین رشد می کنند و در نمایت به گرهها ی برگ می رسد. در درخت تصمیم، هرگره داخلی یا غیر برگ با یک ویژگی مشخص می شود این ویژگی سوالی را در رابطه با مثال ورودی مطرح می کند. در هر گره داخلی به تعداد جوابهای ممکن با این سوال شاخه وجود دارد که هریک با مقدار آن جواب مشخص می شود. برگها ی این درخت با یک کلاس و یا یک دسته از جوابها مشخص می شوند.

علت نامگذاری آن با درخت تصمیم این است که این درخت فرایند تصمیم گیری برای تعیین دسته یک مثال ورودی را نشان می دهد. درخت تصمیم در مسایلی کاربرد دارد که بتوان آنها را به صورتی مطرح نمود که پاسخ واحدی به صورت نام یک دسته یا یک کلاس ارائه دهند. برای مثال می توان درخت تصمیمی ارایه داد که به این سوال پاسخ دهد، بیماری مریض کدام است؟ یا آیا مریض به هپاتیت مبتلا است؟ درخت تصمصم برای مسائلی مناسب است که مثالهای آموزشی بصورت زوج(مقدار-ویژگی) مشخص شده باشند.

ارتباط مستقیمی بین درخت تصمیم و نمایش توابع منطقی وجود دارد درواقع هر درخت تصمیم ترکیب فصلی گزاره های عطفی است.

---

[۱] Recursive partitioning
[۲] Data mining
[۳] Classification tree
[۴] Regression tree



شکل۱. درخت طبقه بندی نتایج موفقیت کاندیداها در سطح بخش ها ی مختلف

یک نمونه از قواعد استخراجی از درخت تصمیم فوق به صورت زیر است:

$$obama\ wins = (county\ black\ populatio > ۲۰٪) \ \vee (county\ black\ population > ۲۰٪ \wedge high\ school\ graduation\ rate > ۷۸٪ \ \wedge high\ school\ graduation\ \ rate > ۸۷\%)$$

II. ابزارها و نحوه کاربرد آنها

در این بخش به نسخه نصبی نرم افزارها و همچنین به چگونگی نصب کردن این نرم افزارها پرداخته شده است.

دراین بخش موارد زیر توضیح داده شده است که عبارتند از :۱-tableau ۲- R packages ۳-R serve ۴- R part ۵- Random Forest

۱- Tableau

برای انجام اعمال داده کاوی مانند درخت تصمیم و یادگیری ماشین باید نسخه ۸.۱ tableau Desktop بر روی ویندوز نصب شود. این نسخه از نرم افزار به صورت مجانی بر روی سایت tableau وجود دارد. مراحل نصب نرم افزار نیز به صورت جزئی در همان سایت نوشته شده است. علت استفاده از این نسخه نسبت به سایر نسخه ها این است که اسکریپت های R در نسخه های دیگر مانند public قابل پشتیبانی نیست. و همچنین ارتباط خوبی که نسخه Desktop با R دارد نیز انگیزه ما را برای استفاده از این نسخه بیشتر کرده است. لازم به ذکر است که این نتایج به صورت تجربی بدست آمده اند هرچند شاید جایی عنوان شده باشد.

۲- R Packages

به وسیله این ابزار مفید می توانید درختان تصمیم بدست آمده به وسیله tableau را در محیط گرافیکی نیز رسم کرده و مشاهده کنید. برای انجام این اعمال آخرین نسخه را از سایت http://Cran.rstudio.com دانلود کرده و نصب کنید. بعد از نصب R حال باید کتابخانه های مربوط به آن را نیز نصب کنید[۵].

۳- R serve

برای نصب کتابخانه R serve در داخل کنسول R دستورات زیر را به ترتیب وارد می کنیم[۶].

```
install. packages("Rserve")
library(Rserve)
Rserve()
```

برای اجرای دوباره R serve فقط دو دستور انتهایی را در کنسولR وارد کنید.

حال برای اینکه بتوانید از اسکریپت های R داخل نرم افزار tableau استفاده کنید باید به محیط tableau رفته و فرمان زیر را اجرا کنید.Help->setting and performance->Manage R connection

پنجره ای مانند پنجره شکل ۲ باز می شود. مانند شکل اطلاعات را وارد کنید. برای اطمینان خاطر از صحیح بودن اتصال روی دکمه Test connection کلیک کنید و در نهایت دکمه OK را فشار دهید تا عملیات تمام شود.

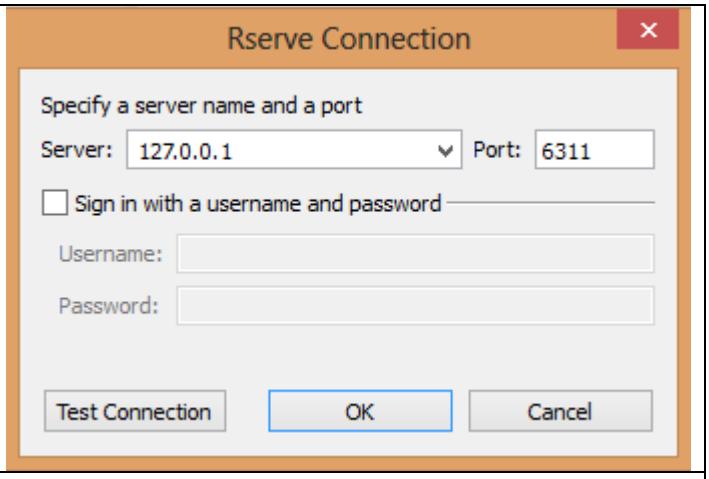
۴- R part

از این کتابخانه به صورت مفید برای ساخت درختهای طبقه بند و رگرسیون استفاده شده است. برای نصب این کتابخانه

---

[۵] این کار بدین دلیل است که از توابع داخل این کتابخانه ها در محیط tableauبرای انجام یادگیریهای بر اساس درخت تصمیم و درخت رگرسیون و شبکه عصبی استفاده می کنیم.

[۶] دقت شود که R نسبت به حروف کوچک و بزرگ حساس است.



شکل۲.پنجره R serve connection در tableau

دستورات زیر را به ترتیب در کنسول R می نویسیم.

```
install. packages("rpart")
library(rpart)
```

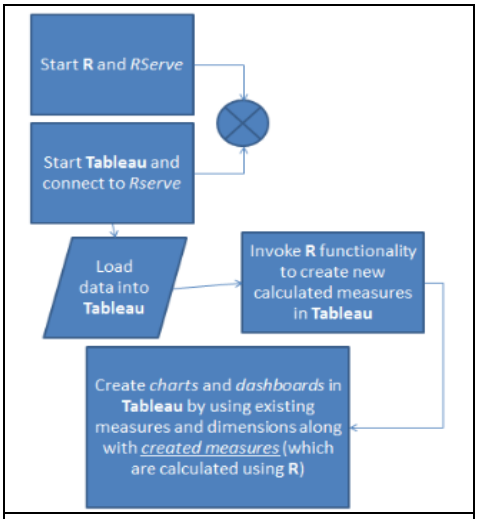
برای استفاده دوباره از R part فقط دستور آخری را در کنسول R تایپ شود. حال می توان از دستورR part در محیط نرم افزار tableau استفاده کرد.

۵- Random Forest

تعداد زیادی روش های یادگیری ماشین وجود دارد که از درخت تصمیم بهره خوب می برند. یکی از شناخته ترین این روش ها کلاس بند Random Forest است که چندین درخت تصمیم را ساخته و خروجی کلاس را می دهد. برای نصب این کتابخانه مفید در R و استفاده از آن در محیط tableau دستورات زیر را به ترتیب در کنسول R وارد کنید.

```
install. packages("randomForest ")
library(randomForest)
```

نمودار دیاگرام شکل ۳ یک دید کلی به چگونگی استفاده از اسکریپت های R در محیط tableau را مثلا برای یک مثال ساده نشان داده است:



شکل۳. دیاگرام کاربرد R در tableau

III. عمل و تئوری ساخت درخت تصمیم با استفاده از rpart و توضیح شبکه عصبی

درختان طبقه بند و رگرسیون را می توان از طریق rpart package تولید کرد.

۱- چگونگی رشد درخت

برای رشد درخت از دستور زیر استفاده می کنیم.

$$rpart(formula, data =, method =, control =)$$

Formula	با این فرمت است: outcome ~ predictor۱+predictor۲+predictor۳+etc
Data	قالب داده تعیین شده
Method	"class" برای درخت طبقه بندی؛"anova" برای درخت رگرسیون
Control	پارامتر انتخابی برای کنترل رشد درخت. به عنوان مثال $control = rpart.control(minsplit = ۳۰, cp = ۰.۰۰۱)$ تاکید می کند که قبل از تقسیم بندی کوچکترین تعداد مشاهدات در یک نود ۳۰ باشد. و آن بخش باید با فاکتور ۰.۰۰۱ مقدار نامناسب بودن درخت را کاهش دهد.(ضریب پیچیدگی هزینه)

۲- بررسی کردن نتایج

توابع زیر جهت بررسی کردن نتایج به ما کمک می کنند.از این دستورات در محیط R استفاده می شود.

load(“”)	در قسمت "" درخت تصمیم ذخیره شده نوشته می شود.
print(fit)	پرینت نتایج
plot(fit)	نمایش درخت تصمیم
text(fit)	ایجاد برچسب نمایش درخت تصمیم

۳- هرس کردن درخت(prune tree)

برای توضیح این بخش ابتدا پدیده Overfitting را بیان کرده و سپس توضیح کامل این دستور در tableau شرح داده می شود.

OVERFITTING

برای فرضیه ای مثل h متعلق به فضای فرضیه H دونوع خطا تعریف می شود:۱- خطا روی داده های آموزشی errortrain(h) و ۲- خطا روی کل داده ها ی errorD(h)،D . می گوییم برای فرضیه $h \in H$ روی داده های آموزشی Overfitting رخ می دهد اگر فرضیه ای مثل $h. \in H$ وجود داشته باشد که:

$$errortrain(h) < errortrain(h.) \ and \ \ errorD(h) > errorD(h.)$$

دلایل بروز Overfitting عبارتند از: ۱- وجود نویز در داده های آموزشی ۲- تعداد کم مثالهای آموزشی. برای پرهیز از Overfitting می توان ۱- جلوگیری از رشد درخت قبل از رسیدن به مرحله ای که بتوان بطورکامل داده های آموزشی را دسته بندی کرد.۲- اجازه به رشد کامل درخت و سپس حرس کردن شاخه هایی که مفید نیستند. در عمل روش دوم بیشتر استفاده شده است زیرا تخمین اندازه صحیح درخت کار ساده ای نیست.

در نرم افزار tableau هرس کردن درخت به اندازه مطلوب با دستور $prune(fit, cp =)$ انجام می شود. اصولا برای بررسی کردن نتایج خطای تصحیح شده از printcp() استفاده می شود. که در این صورت باید پارامتر پیچیدگی مربوط به کوچکترین خطا را انتخاب و در تابع prune() قرار دهید. به صورت ساده تر می توان از قطعه کد زیر استفاده کرد.

$$fit\$cptable[which.\min(fit\$cptable[,\text{xerror}]),\text{CP}]$$

که به صورت اتوماتیک پارامتر پیچیدگی مربوط به کوچکترین خطای تصحیص شده را انتخاب می کند.

۴- Random Forest

Random Forest دقت پیشگویی را با استفاده از تولید تعداد زیادی درخت خودراه انداز شده[۷] بیشتر می کند(بر اساس نمونه های تصادفی از متغیرها)، و یک نمونه را با استفاده از هر درخت در این جنگل[۸] جدید کلاس بندی می کند. و یک نتیجه نهایی پیش بینی شده را با ترکیب نتایج همه درختان تصمیم گبری می کند(یک میانگین در رگرسیون ، و یک رای اصلی در طبقه بندی).نحو صحیح random forest در بخش مثال عملی با Random Forest توضیح داده شده است.

---

۷ Bootstrapped tree

۸ forest

IV. درخت تصمیم

هنگامی که داده ها ویژگی های زیادی دارند و تعامل آنها به صورت غیر خطی است(یعنی به صورت خطی جدایی پذیر نیستند)، آنگاه پیدا کردن یک فرمول رگرسیون کلی مانند فرمول پیش بینی یکه که روی همه دیتاست بکار رود بسیار مشکل است. یک روش جایگزین ، بخش کردن فضا به نواحی کوچکتر است سپس تبدیل آنها به قطعات کوچکتر(قطعه بندی بازگشتی) تا زمانی که همه قطعات بزرگ، قابل شرح به وسیله یک مدل ساده باشند. بنابراین دو نوع اصلی از درخت تصمیم وجود دارد که عبارتند از: ۱- درخت طبقه بندی؛ که نتایج پیش بینی شده ، کلاسی بر طبق داده ها می باشد.۲- درخت رگرسیون؛ که نتایج پیش بینی شده یک متغیر پیوسته است، مانند اعداد اعشاری نظیر قیمت کالا.

در این بخش ابتدا درخت رگرسیون توضیح داده شده و مثالهای عملی با نرم افزار tableau حل شده اند، سپس درخت طبقه بندی نیز با مثالهای عملی در نرم افزار tableau توضیح داده شده است.

۱- درخت رگرسیون

تا اینجا تمام مطالب تئوری و پایه در مورد فهم و پیاده سازی درخت رگرسیون گفته شد. بنابراین در این بخش ساخت درخت رگرسیون در نرم افزار tableau توضیح داده می شود. این بخش بدین صورت توضیح داده شده است که فقط اسم دیتاست را نوشته و برای آن دیتاست دستورات ساخت درخت رگرسیون در نرم افزار tableau عینا عکس گرفته شده و قرار داده شده است. مطالعه جزئیات دیتاست و همچنین توضیحات درباره متغییر های وابسته و مستقل و مفهوم فرضیه را به خود خواننده واگذار می کنیم و فرض بر آن است که خواننده این مباحث آماری را آشنایی دارد. دیتا ست ها از سایت UCI استفاده شده اند.

مثال یک) درخت رگرسیون برای دیتاست CPU Performance

سوال- پیش بینی کارایی CPU (PRP)با توجه به مقادیر MYCT،MMIN،MMAX،CHMIN،CHMAX،CACH را با درخت رگرسیون انجام دهید.

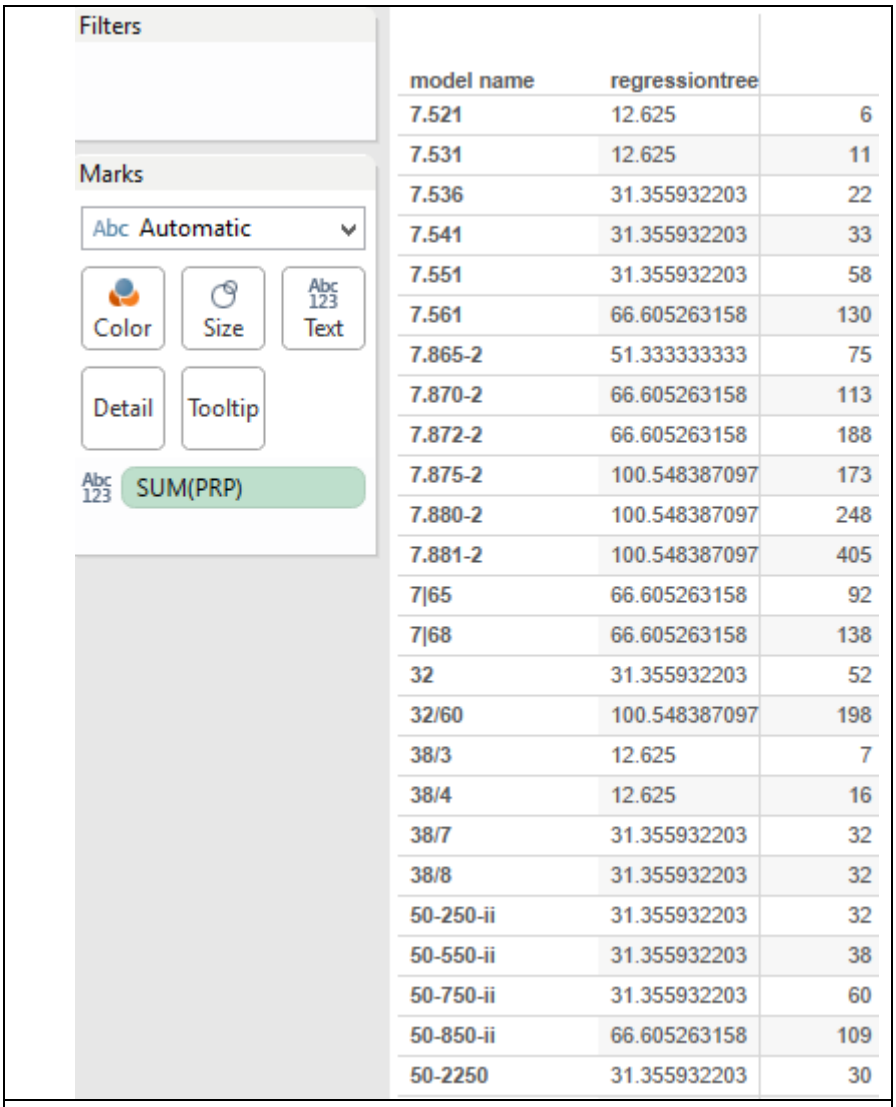
جواب) همانطور که بیان شد نتایج حاصل از درخت رگرسیون یک متغییر اعشاری است بنابرین بارفتن به منوی Analysis و انتخاب گزینه create calculated field قطعه کد شکل ۴ را در پنجره باز شده وارد می کنیم.

```
SCRIPT_REAL('library(rpart);
masregtree = rpart(PRP ~ CACH + CHMAX + CHMIN + MMAX + MMIN + MYCT,
            method="anova",
            data.frame(PRP =factor(.arg1),
                       CACH = factor(.arg2),
                       CHMAX =factor(.arg3),
                       CHMIN =factor(.arg4),
                       MMAX =factor(.arg5),
                       MMIN=factor(.arg6),
                       MYCT=factor(.arg7)));
t(data.frame(predict(prune(masregtree,0.028), type = "vector")))[1,]',
AVG([PRP]),AVG([CACH]),AVG([CHMAX]),AVG([CHMIN]),AVG([MMAX]),AVG([MMIN]), AVG([MYCT]))
```

شکل۴. قطعه کد ساخت درخت رگرسیون

دقت داشته باشید که متغیر های گسسته به عنوان فاکتور در نظر گرفته می شوند. درخت رگرسیون در متغییر masregtree قرار می گیرد.

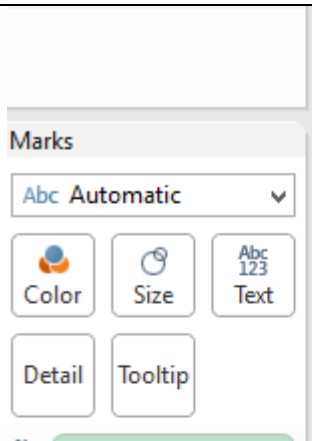
سپس اسمی به قطعه کد فوق داده و دکمه OK را کلیک کنید.قطعه ای از نتایج نشان داده شده به صورت شکل ۵ است.



model name	regressiontree	
7.521	12.625	6
7.531	12.625	11
7.536	31.355932203	22
7.541	31.355932203	33
7.551	31.355932203	58
7.561	66.605263158	130
7.865-2	51.333333333	75
7.870-2	66.605263158	113
7.872-2	66.605263158	188
7.875-2	100.548387097	173
7.880-2	100.548387097	248
7.881-2	100.548387097	405
7|65	66.605263158	92
7|68	66.605263158	138
32	31.355932203	52
32/60	100.548387097	198
38/3	12.625	7
38/4	12.625	16
38/7	31.355932203	32
38/8	31.355932203	32
50-250-ii	31.355932203	32
50-550-ii	31.355932203	38
50-750-ii	31.355932203	60
50-850-ii	66.605263158	109
50-2250	31.355932203	30

شکل ۵- بخشی از نتایج اعمال درخت رگرسیون روی داده های CPU Performance

برای همین دیتاست موفق شدیم عملیات رگرسیون خطی را نیز بدست آوریم در سند قبلی چگونگی این عملیات توضیح داده شده است. لذا در شکل ۶ بخشی از نتایج درخت رگرسیون در مقابل رگرسیون خطی نشان داده شده است.



model name	regressiontree	ERP	
7.521	12.625	19	6
7.531	12.625	22	11
7.536	31.355932203	31	22
7.541	31.355932203	41	33
7.551	31.355932203	47	58
7.561	66.605263158	99	130
7.865-2	51.333333333	67	75
7.870-2	66.605263158	81	113
7.872-2	66.605263158	149	188
7.875-2	100.548387097	183	173
7.880-2	100.548387097	275	248
7.881-2	100.548387097	382	405
7|65	66.605263158	70	92
7|68	66.605263158	117	138
32	31.355932203	41	52
32/60	100.548387097	199	198
38/3	12.625	18	7
38/4	12.625	20	16
38/7	31.355932203	28	32
38/8	31.355932203	45	32
50-250-ii	31.355932203	20	32
50-550-ii	31.355932203	29	38
50-750-ii	31.355932203	43	60
50-850-ii	66.605263158	53	109
50-2250	31.355932203	25	30

شکل۶- مقایسه درخت رگرسیون و رگرسیون خطی

مثال دو) درخت رگرسیون برای دیتاست Auto MPG

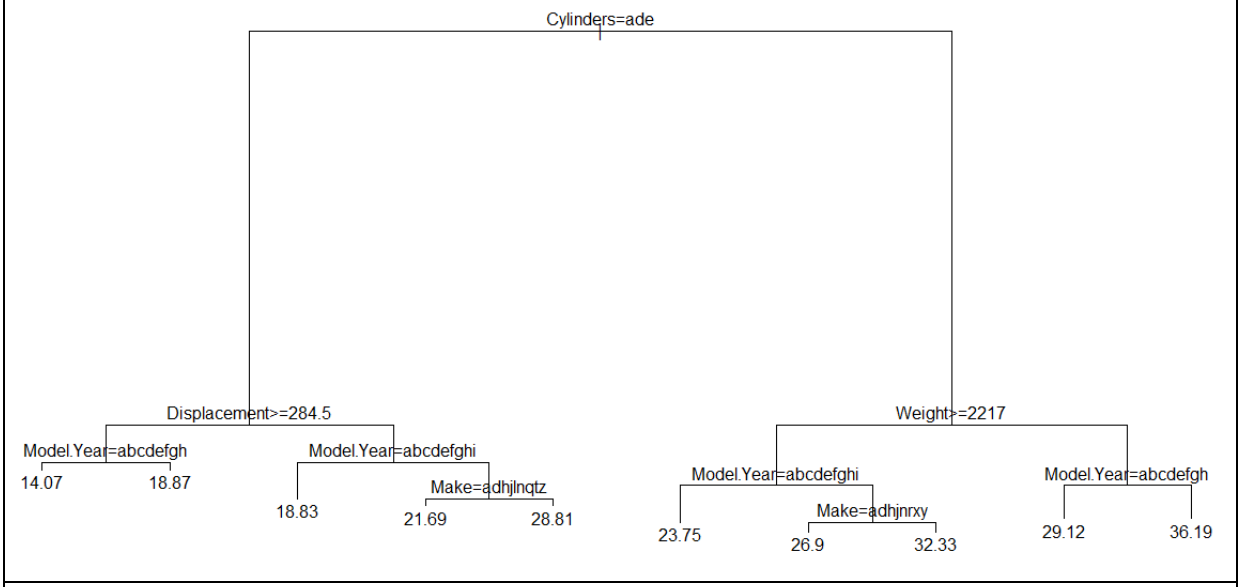
سوال- پیش بینی کنید که مسافت طی شده با یک گالن بنزین برای هر ماشین چقدر است؟ برای اینکار از پیشگو های مساله استفاده کرده و درخت رگرسیونی که tableau برای داده ها بدست می آورد را با R رسم کنید. بخشی از نتایج را نیز در شکل بیاورید.

جواب- اول اینکه منظور از پیشگو ها همان ویژگی های مساله هستند. قطعه کد لازم برای پیش بینی مانند شکل ۷ است.

```
SCRIPT_REAL('library(rpart);
fit = rpart(Mpg ~ Acceleration + Cylinders + Displacement + Horsepower + Make + Model.Year+ Origin + Weight,
method="anova",
data.frame(Mpg = .arg1,
Acceleration =   .arg2,
Cylinders=factor(.arg3), Displacement =.arg4, Horsepower =.arg5, Make=factor(.arg6), Model.Year = factor(.arg7), Origin = factor(.arg8), Weight = .arg9));
t(data.frame(predict(prune(fit,0.028), type = "vector")))[1,]',
AVG([Mpg]),AVG([Acceleration]),AVG([Cylinders]),AVG([Displacement]),AVG([Horsepower]),ATTR([Make]), AVG([Model Year]),ATTR([Origin]),AVG([Weight]))
```

شکل۷- قطعه کد تولید درخت رگرسیون برای دیتاست Auto MPG

درخت رگرسیون تولید شده را با R به صورت شکل ۸ بدست آوردیم. این درخت حاصل کد فوق است که در محیط R رسم شده است.



شکل۸- درخت رگرسیون برای مثال فوق؛ رسم شده در محیطR

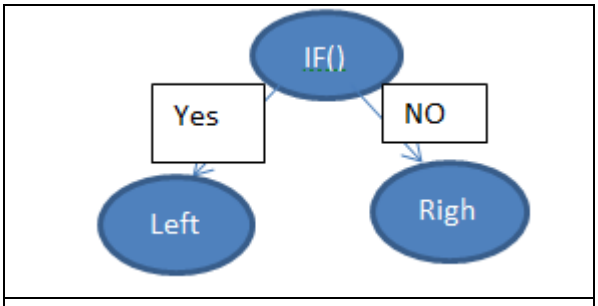
نتایجی که با درخت فوق بدست آمده اند را شکل ۹ نشان داده شده است.این نتایج را نرم افزار tableau می دهد.

CarID	RegressionTr..	SUM(Mpg)
1	14.764948454	18.00
2	14.764948454	15.00
3	14.764948454	18.00
4	14.764948454	16.00
5	14.764948454	17.00
6	14.764948454	15.00
7	14.764948454	14.00
8	14.764948454	14.00
9	14.764948454	14.00
10	14.764948454	15.00
11	14.764948454	15.00
12	14.764948454	14.00
13	14.764948454	15.00
14	14.764948454	14.00
15	20.010752688	22.00
16	20.010752688	18.00
17	20.010752688	21.00
18	20.010752688	21.00
19	14.764948454	10.00
20	14.764948454	10.00
21	14.764948454	11.00
22	23.7515625	28.00
23	29.119565217	25.00
24	20.010752688	19.00

شکل۹- نتایج اعمال روی دیتاست در محیط tableau

۲- درخت طبقه بندی

همانطور که گفته شد برای درخت طبقه بند، نتایج پیش بینی شده یک کلاس است. که این کلاس بندی طبق داده ها می باشد. بنابراین درخت طبقه بندی را برای دیتاست هایی بکار می بریم که داده ها در آن قابلیت کلاس بندی را داشته باشند. دقت شود که نحوه خواندن درختان تصمیم به صورت شکل ۱۰ است.



شکل ۱۰- نحوه پیمایش درختان تصمیم

برای توضیح دو دیتاست بکار برده ایم. برای این دو دیتاست سوالات زیادی مطرح می کنیم. و پاسخ کامل آنها را نیز شرح می دهیم.

مثال یک) درخت طبقه بندی برای دیتاست CPU Performance

سوال- درخت طبقه بندی را با استفاده از نرم افزار tableau جهت پیش بینی کلاس هر مدل CPU بدست بیاورید. از مقدار PRPبرای کلاس بندی داده های دیتاست استفاده کنید، جدول۳.

جدول۳. کلاس بندی داده های دیتاست

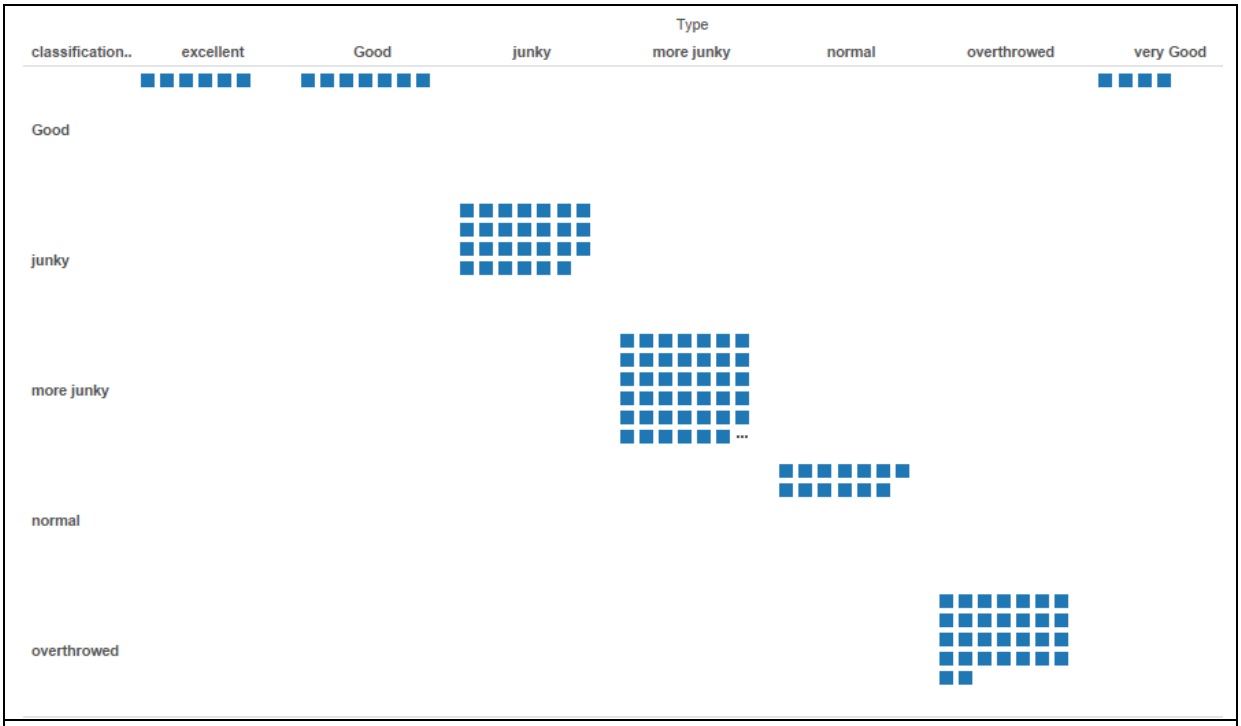
محدوده	نام کلاس
۰ تا ۲۰	منقرض شده(overthrow)
۲۱ تا ۱۰۰	خیلی بد(more junky)
۱۰۱ تا ۲۰۰	بد(junky)
۲۰۱ تا ۳۰۰	متوسط(normal)
۳۰۱ تا ۴۰۰	خوب(Good)
۴۰۱ تا ۵۰۰	خیلی خوب(very Good)
بالای ۵۰۰	عالی(excellent)

جواب) قطعه کد پیش بینی کلاس ها در شکل ۱۱ آمده است. در این قطعه کد نحوه ذخیره کردن درخت تصمیم نیز نشان داده شده است.

```
SCRIPT_STR('library(rpart);mastree = rpart(Type ~ CACH+CHMAX+CHMIN+MMAX+MMIN+MYCT+PRP,
method="class",
data.frame(Type = .arg1,
           CACH = .arg2,
           CHMAX= .arg3,
           CHMIN= .arg4,
           MMAX= .arg5,
           MMIN= .arg6,
           MYCT= .arg7,PRP = .arg8));save(mastree,file="D:/mohammadi/clssificationtree.rda");
mastree<-prune(mastree,fit$cptable[which.min(fit$cptable[,"xerror"]),"CP"]);
t(data.frame(predict(mastree, type = "class")))[1,]',
ATTR([Type]),AVG( [CACH]),AVG([CHMAX] ),AVG([CHMIN] ),AVG([MMAX] ),AVG([MMIN] ),AVG([MYCT]),AVG([PRP] ))
```

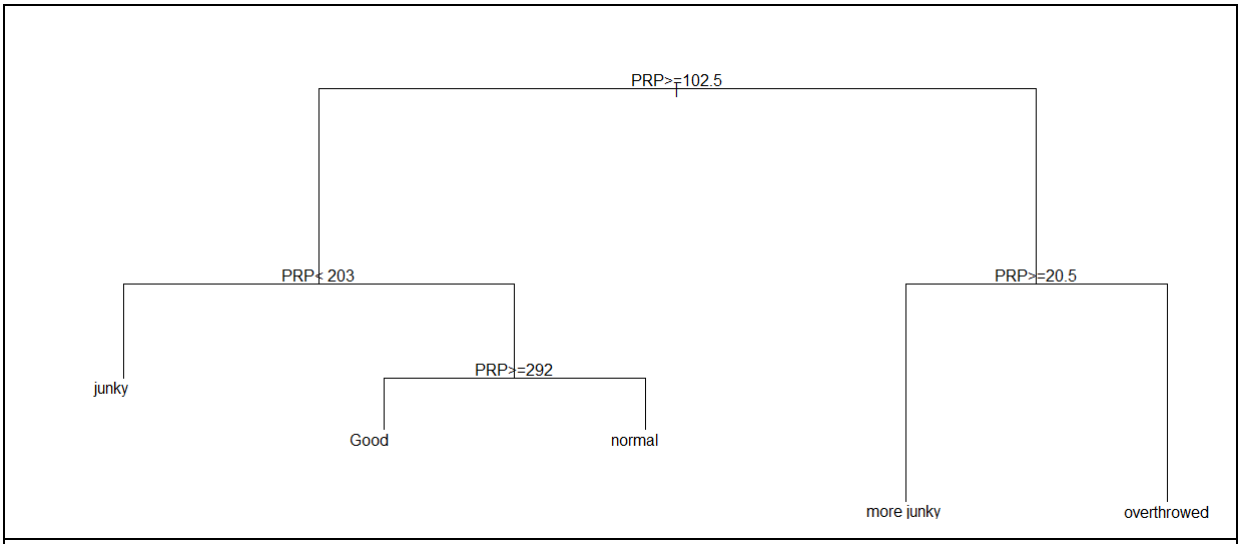
شکل۱۱- قطعه کد جهت بدست آوردن درخت طبقه بندی در tableau

نتایج حاصل از این درخت بر روی داده های دیتاست به صورت شکل ۱۲ است. این خروجی را tableau می دهد.



شکل۱۲- قرار گیری نمونه ها در کلاسها با توجه به درخت قطعه بند

حال درخت تصمیمی که tableau با استفاده از آن کلاس بندی های فوق را انجام می دهد و آن را ذخیره کردیم در محیط R رسم می کنیم، این درخت در شکل ۱۳ آمده است. دستورات فراخوانی درخت در کنسول R در شکل ۱۴ آمده است.



شکل۱۳- درخت تصمیم تولید شده برای انجام عملیات کلاس بندی توسط tableau در محیط R

```
> load("D:/mohammadi/clssificationtree.rda")
> plot(mastree)
> text(mastree)
```

شکل ۱۴- دستورات لازم برای فراخوانی درخت در کنسول R

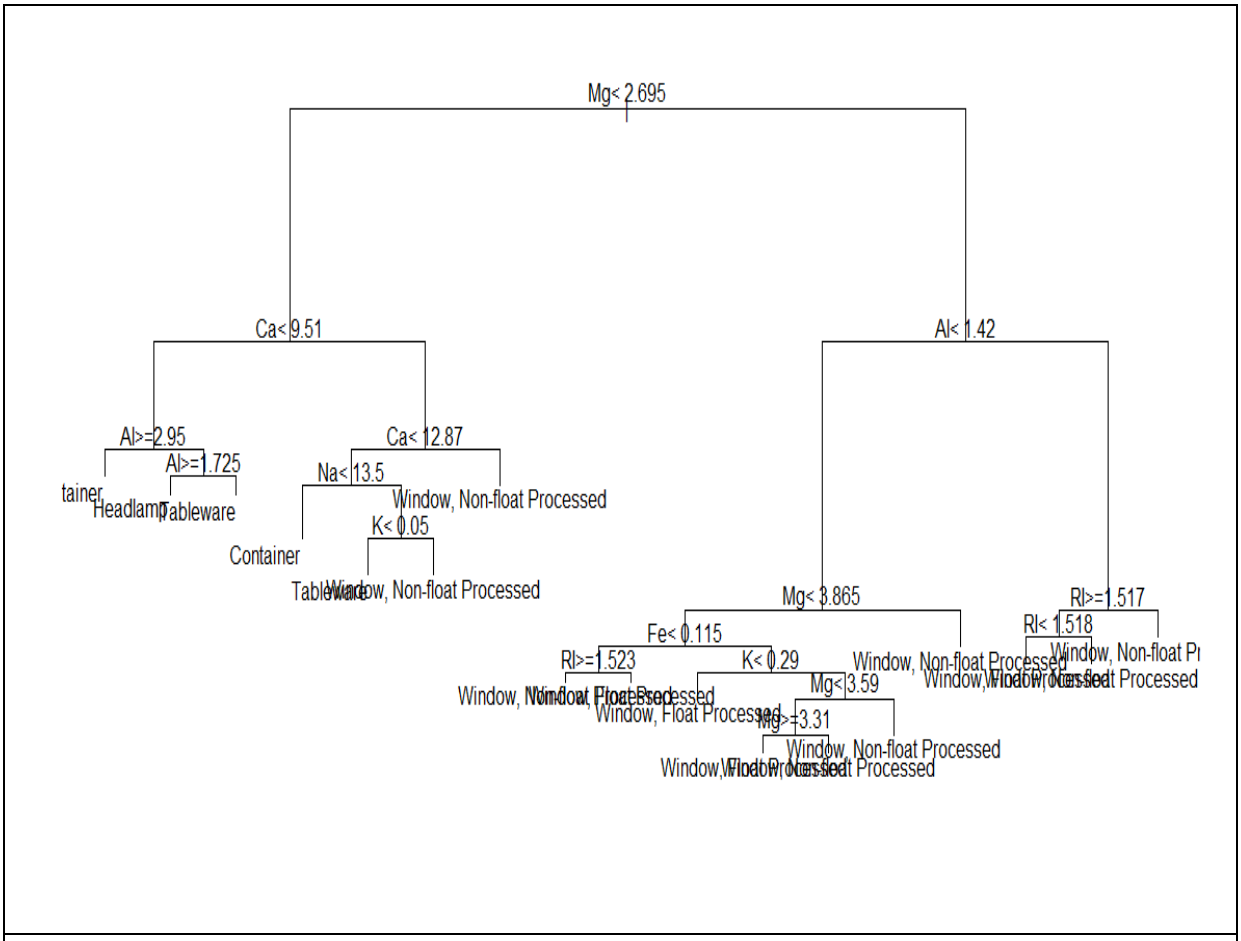
بسیاری از عملیات دیگر در فایل ارسالی انجام شده است که به دلیل زیاد شدن حجم فایل متنی از توضیح آنها خواهم گذشت و دیتاست دیگری را در مثال بعدی برای همین عملیات شرح می دهم.

پرسش۱- مثال فوق کلاس بندی بدون کنترل انجام شده است. کلاس بندی با کنترل آن را انجام دهید و همچنین بررسی کنید که به ازای چه مقادیری از minsplit و cp معادل کلاس بندی بدون کنترل حاصل می شود؟

مثال دو) درخت طبقه بند برای دیتاست Glass

سوال- فرض کنید در یک صحنه جنایت شیشه ای شکسته است. حال می خواهیم بررسی کنیم که این شیشه شکسته جزو کدام دسته از شیشه ها می باشد؟ (به منظور جلوگیری از حجیم شدن فایل فقط درخت تصمیم رسم می شود)

درخت تصمیم این پژوهش در شکل ۱۵ نشان داده شده است.



شکل ۱۵- درخت طبقه بند حاصل از دادها با استفاده از R

V. شبکه عصبی random Forest

در مورد ابن امکان قبلا توضیح داده شد. بنابراین نحوه پیاده سازی را در tableau بیان می کنیم. برای توضیح این بخش به توضیح یک مثال می پردازیم.

اساس کار این شبکه عصبی کلاس بندی بر مبتای درخت طبقه بند است.

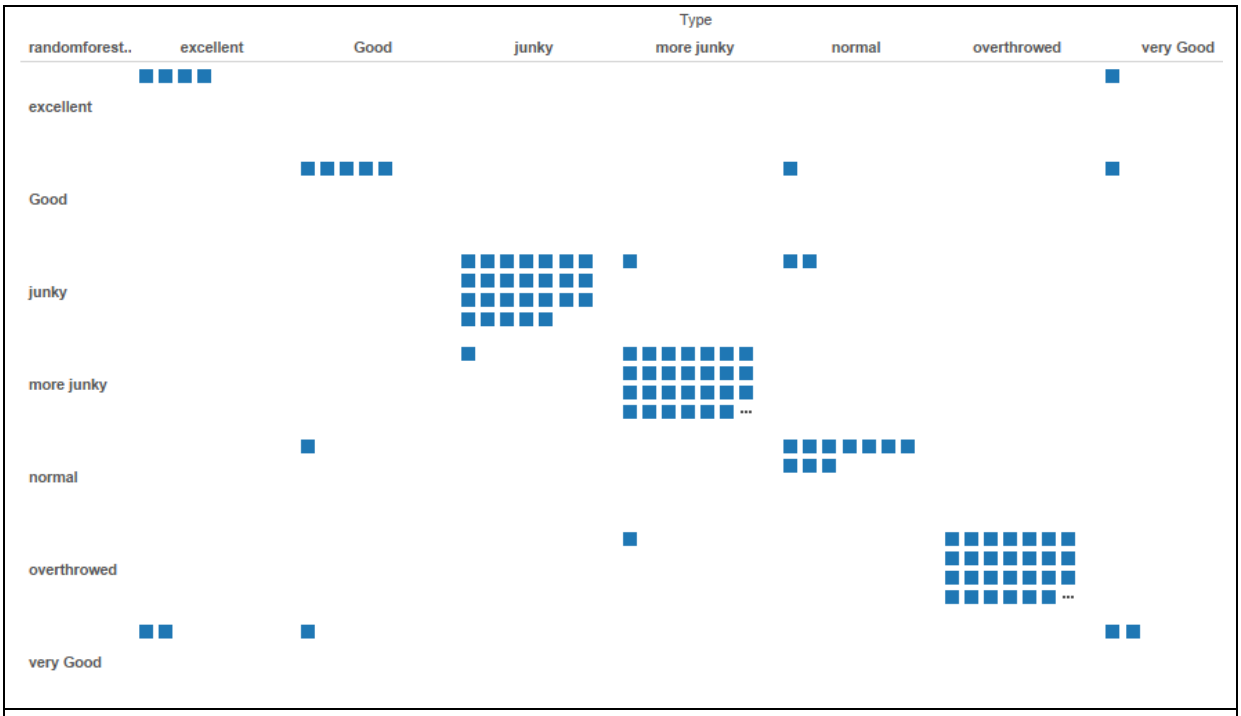
سوال) شبکه عصبی random Forest را برای کلاس بند دیتاست CPU Performance بکار ببرید.

جواب) قطعه کد شکل ۱۶ عملیات random Forest را در tableau نشان داده است.

```
SCRIPT_STR('library(rpart);mastree = randomForest(Type ~ CACH+CHMAX+CHMIN+MMAX+MMIN+MYCT+PRP,
data.frame(Type = .arg1,
           CACH = .arg2,
           CHMAX= .arg3,
           CHMIN= .arg4,
           MMAX= .arg5,
           MMIN= .arg6,
           MYCT= .arg7,PRP = .arg8));
io<-predict(mastree, type = "prob");
colnames(io)[apply(io,1,which.max)]',
ATTR([Type]),AVG( [CACH]),AVG([CHMAX] ),AVG([CHMIN] ),AVG([MMAX] ),AVG([MMIN] ),AVG([MYCT]),AVG([PRP] ))
```

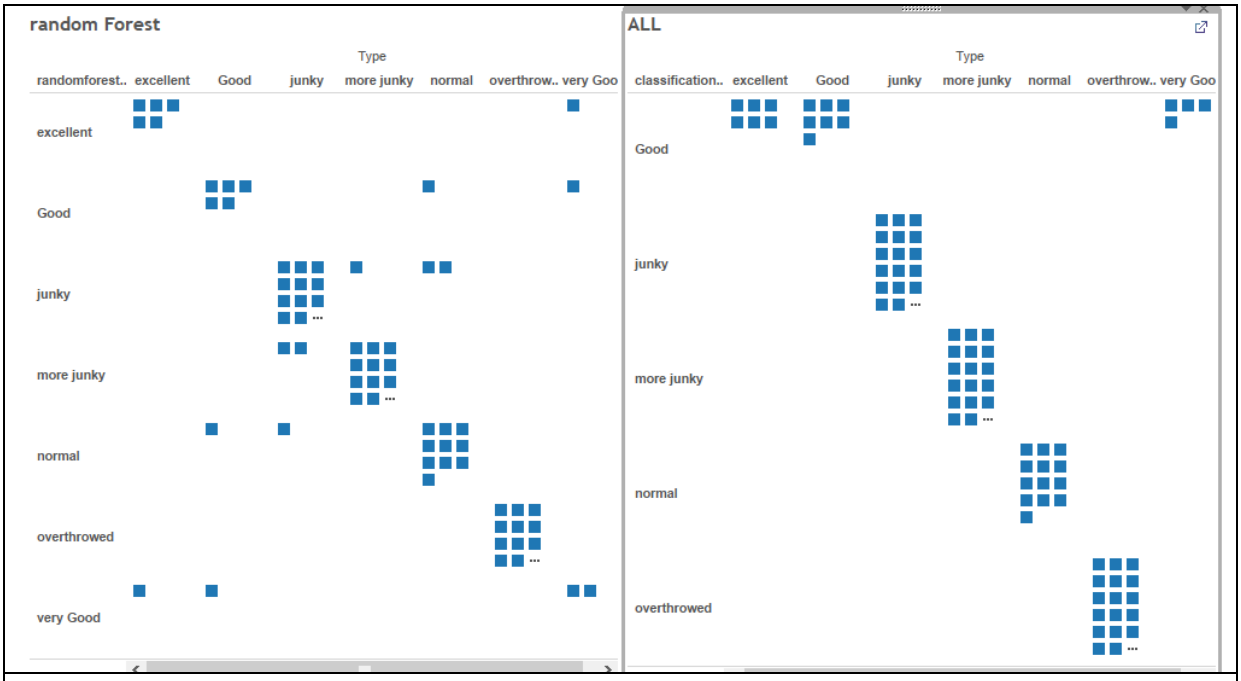
شکل ۱۶- دستورات random Forest در tableau

بعد از آموزش نتایج کلاس بندی به صورت شکل ۱۷ می باشد.



شکل۱۷- کلاس بندی داده ها با شبکه عصبی random Forest

همانند درخت طبقه بند شبکه عصبی random Forest نیز خطا دارد. داده های بد کلاس بندی شده در شکل واضح اند. مقایسه این دو کلاس بند در شکل ۱۸ نشان داده شده است.



شکل ۱۸. مقایسه دو روش درخت طبقه بند و شبکه عصبی random Forest جهت کلاس بندی

شکل ۱۸ نشان می دهد که درخت طبقه بند و شبکه عصبی random Forest برای کلاس بندی هرکدام به ترتیب ۱۰ خطا و ۱۲خطا دارند.

پرسش۲- در پژوهش دیتا ست CPU Performanceنرخ موفقیت درخت طبقه بند چند است؟ (با فرمولی در tableau بدست آورید)

پرسش۳- در پژوهش دیتا ست CPU Performance نرخ موفقیت شبکه عصبی random Forest چند است؟(با فرمولی در tableau بدست آورید)