**گزارش نتایج دادگان درس داده کاوی**

اعضای گروه: پارمیدا خانی، علی بهلولی، هادی مرادی

**مقدمه:**

ما برای دسته بندی داده ها از scikit\_learn و cost complexity pruning برای ساخت classification tree استفاده کردیم. تا پیش بینی کنیم که دلبستگی مراقبین جانبازان از نوع ایمن یا نا ایمن و یا دوسوگرا است. بعد از هندل کردن missing data ها داده ها را به فرمت مناسب برای classification tree در آوردیم.

**شرح دسته بندی:**

برای این کار داده را به دو بخش تقسیم کردیم: 1) ویژگی هایی از داده که برای دسته بندی استفاده می شود 2) ویژگی دلبستگی که قصد پیش بینی آن را داریم.

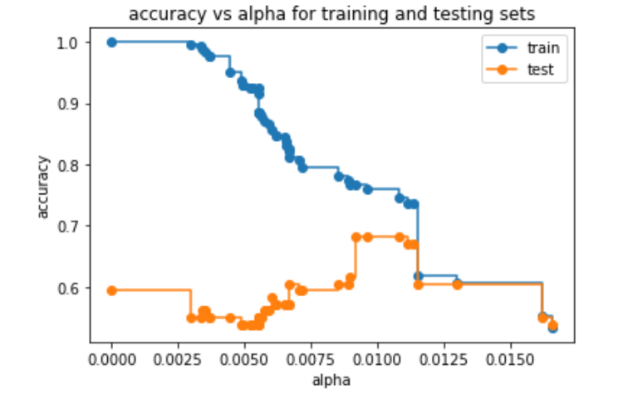
به دلیل اینکه scikit\_learn decision trees داده های پیوسته را پوشش می دهد و داده های categorical را پوشش نمی دهد از one hot encoding استفاده کردیم تا داده های categorical را به صورت چند ستون از داده های باینری درآوریم.

برای one hot encoding از تابع get\_dummies که از کتاب خانه ی pandas می باشد استفاده کردیم. این روش داده هایی که بیشتر از دو دسته دارند را تغییر می دهد بنابراین داده هایی همانند جنسیت و یا گرایش که به دو دسته تقسیم می شوند را بدون تغییر باقی می گذارد. برای ساخت درخت داده را به دو مجموعه ی train و تست تقسیم می کنیم.

با استفاده از امکاناتی که کتابخانه ی sklearn.tree در اختیارمان قرار می دهد شکل گرافیکی decision tree را به دست آوردیم که در ابتدا درخت بسیار بزرگ بود که تصمیم بر هرس کردن آن گرفتیم زیرا در این حالت overfit بر روی داده های train زیاد است پس از cost complexity pruning برای کوچک کردن درخت و همچنین بالا بردن صحت در داده های تست بهره بردیم. برای هرس کردن پارامتر آلفا پارامتر مهمی است که میزان هرس را مشخص می کند به طوری که با افزایش آلفا هرس بیشتری خواهیم داشت.

برای یافتن بهترین مقدار آلفا از نمودار صحت درخت بر اساس مقادیر مختلف آلفا استفاده کردیم که این ارزیابی را بر هر دو داده ی تست و train انجام دادیم.

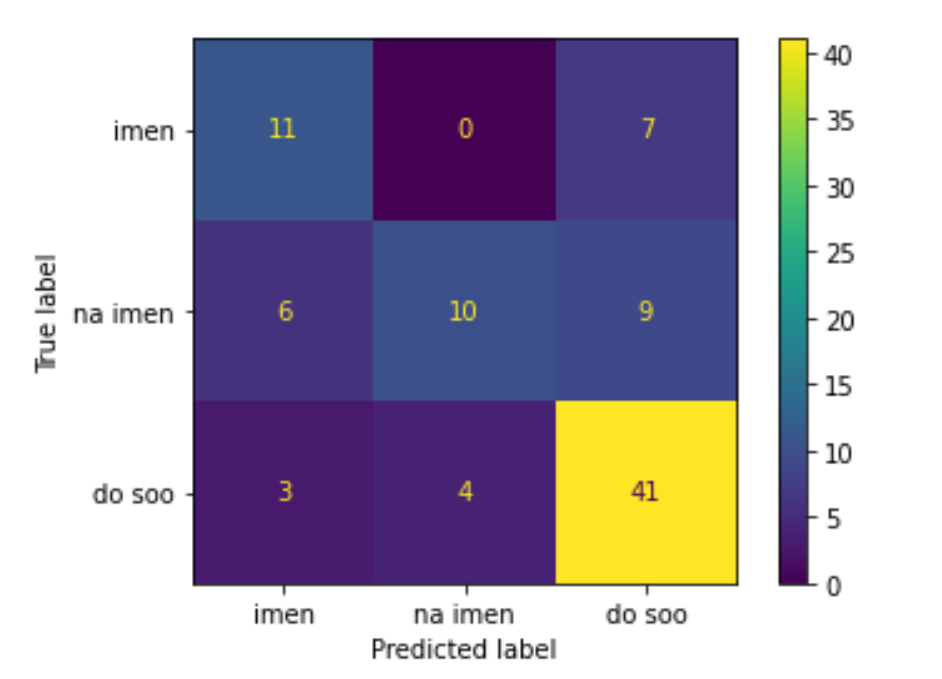
در نمودار حاصل در جایی که داده ی تست بیشترین صحت را دارد و داده ی train از آن نقطه به بعد افت زیادی می کند مقدار آلفا را ذخیره می کنیم.



حال با استفاده از آلفا درختمان را هرس کردیم و confusion matrix در مقایسه با درخت پیش از هرس بیش بینی بهتری برای یافتن سه دسته ی مد نظرمان به ما داد.

در درخت حاصل معیار تقسیم در سطر اول آمده که اگر مقدار آن true باشد وارد شاخه ی سمت چپ می شود و اگر false باشد شاخه ی سمت راست. همچنین مقدار ضریب جینی برای هر نود مشخص شده که هرچه کمتر باشد خلوص آن نود بیشتر است.

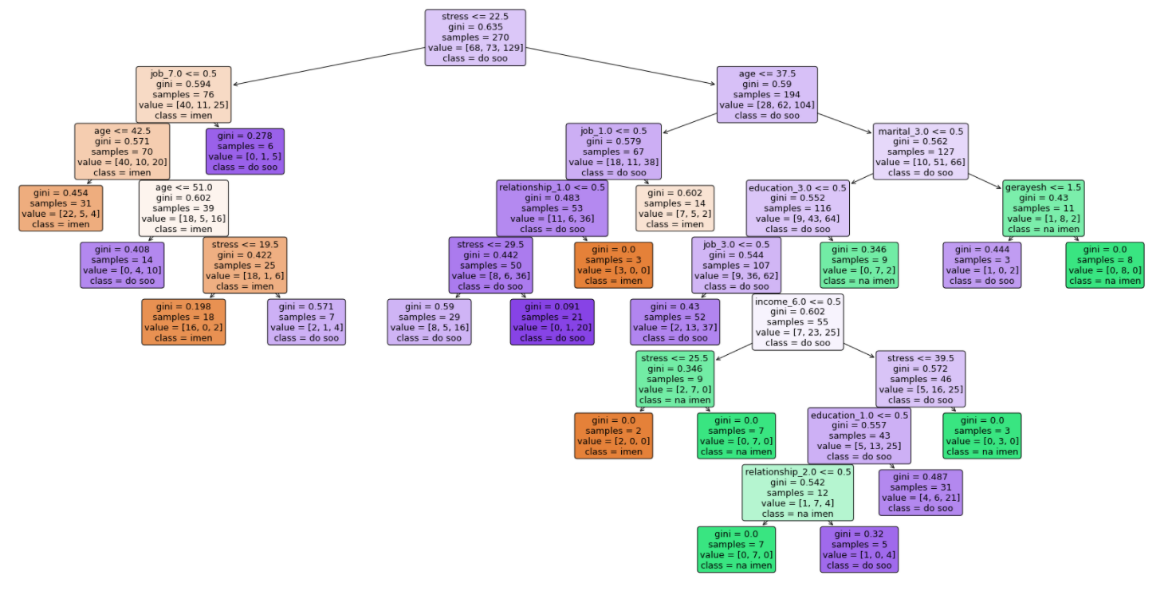
**Confusion matrix** :



**تحلیل decision tree :**

با استفاده از این درخت می توان برای داده های جدید نوع دلبستگی فرد را پیش بینی کرد. همچنین با تحلیل نود هایی که خلوص بیشتری دارند می توان برای نمونه نتایج زیر را گرفت:

1. افرادی که استرس بیشتر از 22 دارند و سن شان کمتر از 38 است و کارمند نیستند و همسر ندارند به احتمال زیاد دلبستگی دوسوگرا دارند.
2. افرادی که استرس کمتر از 20 دارند و شغلی غیر از آزاد دارند و سن شان از 51 بیشتر است به احتمال زیاد دلبستگی ایمن دارند.
3. افرادی که استرس بیشتر از 22 دارند و سن شان از 37 بیشتر است و همسر خود را از دست نداده اند و گرایششان جمع گرایی است به احتمال زیاد دلبستگی ناایمن دارند.



نتیجه گیری:

در بررسی کلی فرهنگ فردگرایی و جمع­گرایی با ابعاد دلبستگی، بین بعد فردگرایی و جمع­گرایی و دلبستگی ایمن در گروه سنی جوانان رابطه آماری معناداری وجود دارد. بین فردگرایی و دلبستگی ایمن ارتباط مستقیم وجود دارد. یعنی با افزایش فردگرایی دلبستگی ایمن نیز بیشتر می­شود در حالی که جمع­گرایی با دلبستگی ایمن رابطه معکوس دارد به این معنی که با افزایش فرهنگ جمع­گرایی در جوانان دلبستگی ایمن کمتر می­شود. هم­چنین بین فرهنگ فردگرایی و دلبستگی دوسوگرا در گروه جوانان رابطه معنادار و مستقیم وجود دارد. با افزایش فرهنگ فردگرایی در گروه جوانان دلبستگی دوسوگرا نیز افزایش می­یابد.

ارتباط معنادار و معکوس بین فرهنگ جمع­ گرایی و دلبستگی نا ایمن دوسوگرا در گروه بزرگسالان وجود دارد. یعنی با افزایش فرهنگ جمع­گرایی دلبستگی نا ایمن دوسوگرا کاهش می­یابد. هم­چنین بین فرهنگ فردگرایی و دلبستگی نا ایمن اجتنابی ارتباط معنادار و معکوس وجود دارد. با افزایش فرهنگ فردگرایی در بالغین میزان دلبستگی اجتنابی نیز افزایش می­یابد. بین ابعاد فرهنگ و ابعاد دلبستگی ارتباط معناداری وجود ندارد.