به نام خدا



دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

درس یادگیری ماشین

استاد ناظرفرد

تمرین دوم

علیرضا مازوچی

۴۰۰۱۳۱۰۷۵

بخش اول: پرسش‌های تشریحی

سوال ۱

الف) اگر مقدار باشد، وقتی که داده validation هر یک از چهار داده کلاس +‌ باشد، مدل به اشتباه آن را – پیش‌بینی می‌کند چراکه داده – در میان آن‌ها فاصله کمتری تا هر چهار داده دیگر نسبت به همدیگر دارد ( فاصله مرکز مربع تا رئوس کمتر از فاصله دو راس مربع است. بدهی است که خود این داده منفی هم به اشتباه دسته‌بندی خواهد شد. داده‌ها منفی طرف دیگر به وضوح درست دسته‌بندی می‌شوند.

باتوجه به اینکه k نباید مقدار زوج داشته باشد، حال باید را بررسی کنیم. در این حالت اتفاقی که برای چهار داده + رخ داد دیگر اتفاق نمی‌افتد چون به هر داده + دو داده + و یک داده – نزدیک است اما همچنان داده – میان آن‌ها به اشتباه دسته‌بندی می‌شود. داده‌های منفی طرف دیگر به وضوح درست دسته‌بندی می‌شوند.

اگر باشد، برای هر یک از داده‌های + همچنان سه داده + نزدیک‌تر وجود دارد که در نتیجه این‌ها به درستی ارزیابی می‌شوند. برای داده‌های – در طرف چپ هم به وضوح نتیجه دسته‌بندی درست خواهد بود. اما همچنان داده – میان داده‌های مثبت به چهار داده + نزدیک است و دسته‌بندی اشتباهی خواهد داشت.

اگر باشد در این صورت برای هر یک از داده‌های + تنها سه داده + باقی می‌ماند و درنتیجه همگی به اشتباه به کلاس – تعلق می‌گیرند. در نتیجه متوجه می‌‌شویم برای این مسئله حالت و با یک خطا بهترین حالت است.

ب)‌ مانند تعیین مقدار هر ابرپارامتر دیگر می‌توان از یک مجموعه validation برای تعیین آن استفاده کرد. یعنی آنکه به ازای مقادیر k مختلف بررسی می‌کنیم که خطای مجموعه validation به چه شکلی بوده است. جایی که کمترین خطا را برای این مجموعه داشتیم احتمالا k مناسب داشته است.

سوال ۲

الگوریتم KNN‌ را می‌توان یک الگوریتم تمایز‌گر دانست؛ چراکه در این الگوریتم حساب می‌شود که یک داده به کدام کلاس تعلق دارد و مرز‌های کلاس‌ها مختلف تشکیل می‌شود (مانند سلول‌های voronoi برای 1NN) و نمی‌توان الگوی یک کلاس را محاسبه کرد و توسط آن داده‌های جدید ایجاد کرد.

الگوریتم درخت تصمیم را هم باید یک الگوریتم تمایزگر دانست؛ چراکه در این الگوریتم هم باز صرفا بررسی می‌شود که یک داده به کدام کلاس تعلق دارد و می‌توان مرز‌ها را برای کلاس‌های مختلف تشکیل داد ولی نمی‌توان توزیع هر کلاس را بدست آورد.[[1]](#footnote-1)[[2]](#footnote-2)

سوال ۳

الف)

ب)

????

ج)

???

د)؟

سوال ۴

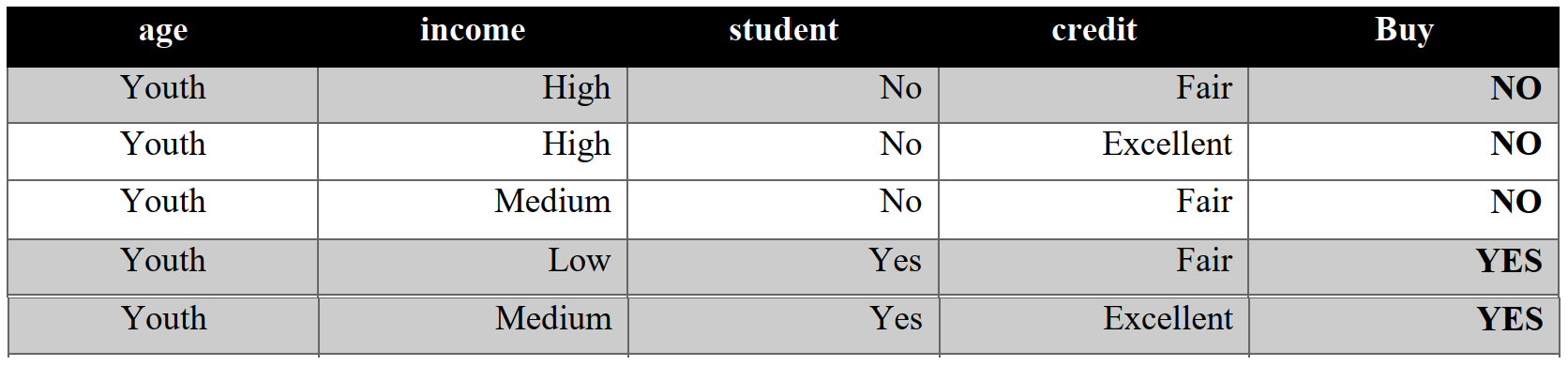
الف)

ب)

ابتدا باید بررسی کرد که برای ریشه درخت کدام ویژگی مناسب‌تر است.

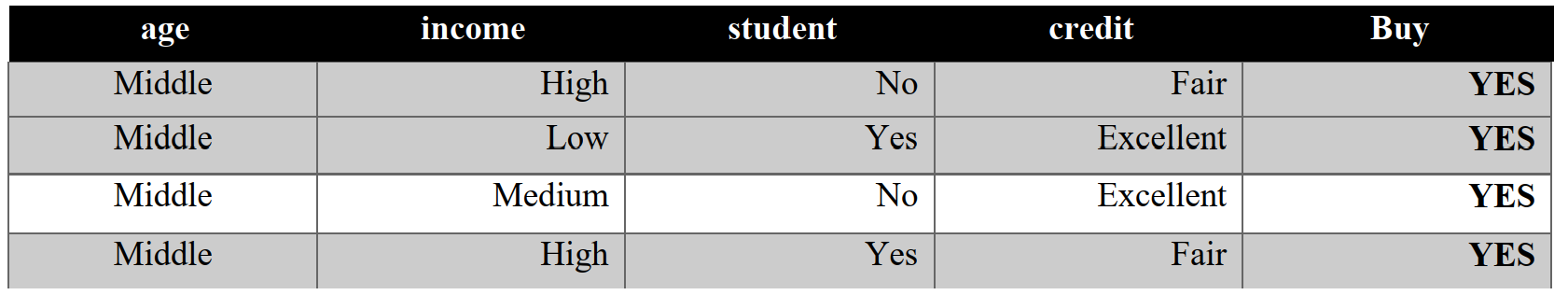
به نظر می‌رسد برای ریشه پرسش راجع به age بهترین گزینه است.

حال باید بررسی کرد که برای شاخه age = youth کدام ویژگی مناسب‌تر است.

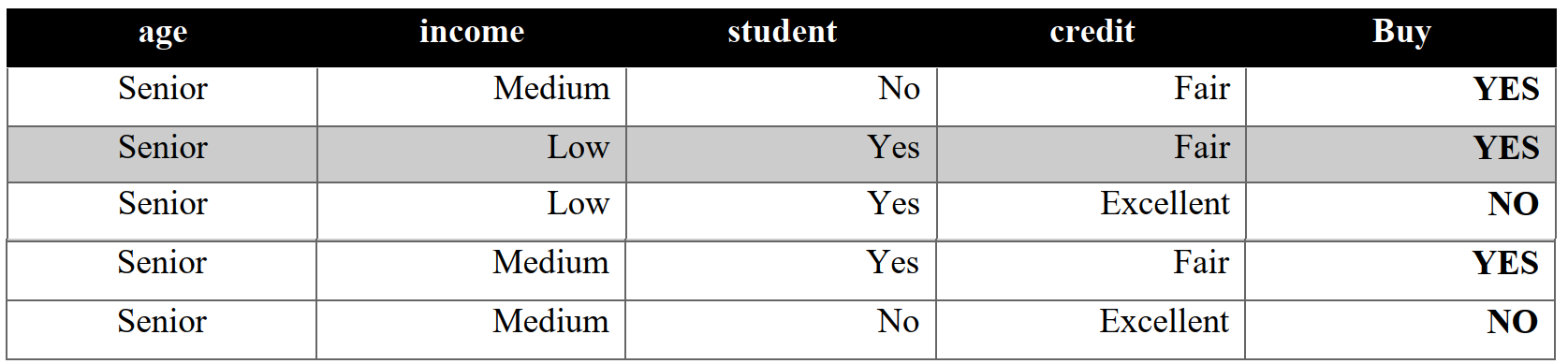


طبیعتا برای این شاخه هم می‌توان مشابه با محاسبات گره ریشه پیش رفت؛ اما پیش از آن توجه کنید که اگر student‌ مورد سوال قرار بگیرد قطعا دو گره خالص ایجاد خواهد شد این درحالی است که در مابقی ویژگی‌ها حداقل یک گره ناخالص باقی می‌ماند. ( برای income، مقدار medium؛ برای credit، مقدار fair) پس قطعا Gain ویژگی Student برابر بیشینه مقدار که انتروپی همین گره age= youth باشد است که از مابقی بیشتر است و نیازی به محاسبات بیشتر نیست.

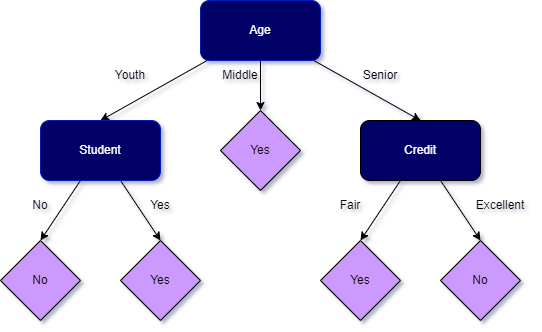
برای شاخه age = middle یک گره خالص باقی می‌ماند که نیاز به ادامه ندارد و می‌دانیم به پاسخ yes خواهیم رسید.



برای شاخه آخر یعنی age = senior، واضحا و مشابه با بحث گره age =youth انتخاب ویژگی credit بهترین گزینه است چراکه دو گره خالص حاصل می‌شود و Gain این ویژگی برابر با آنتروپی گره age = senior‌ است که بیشترین Gain ممکن است. برای دو ویژگی دیگر هم مشخص است که گره‌های حاصل تماما خالص نخواهند بود.

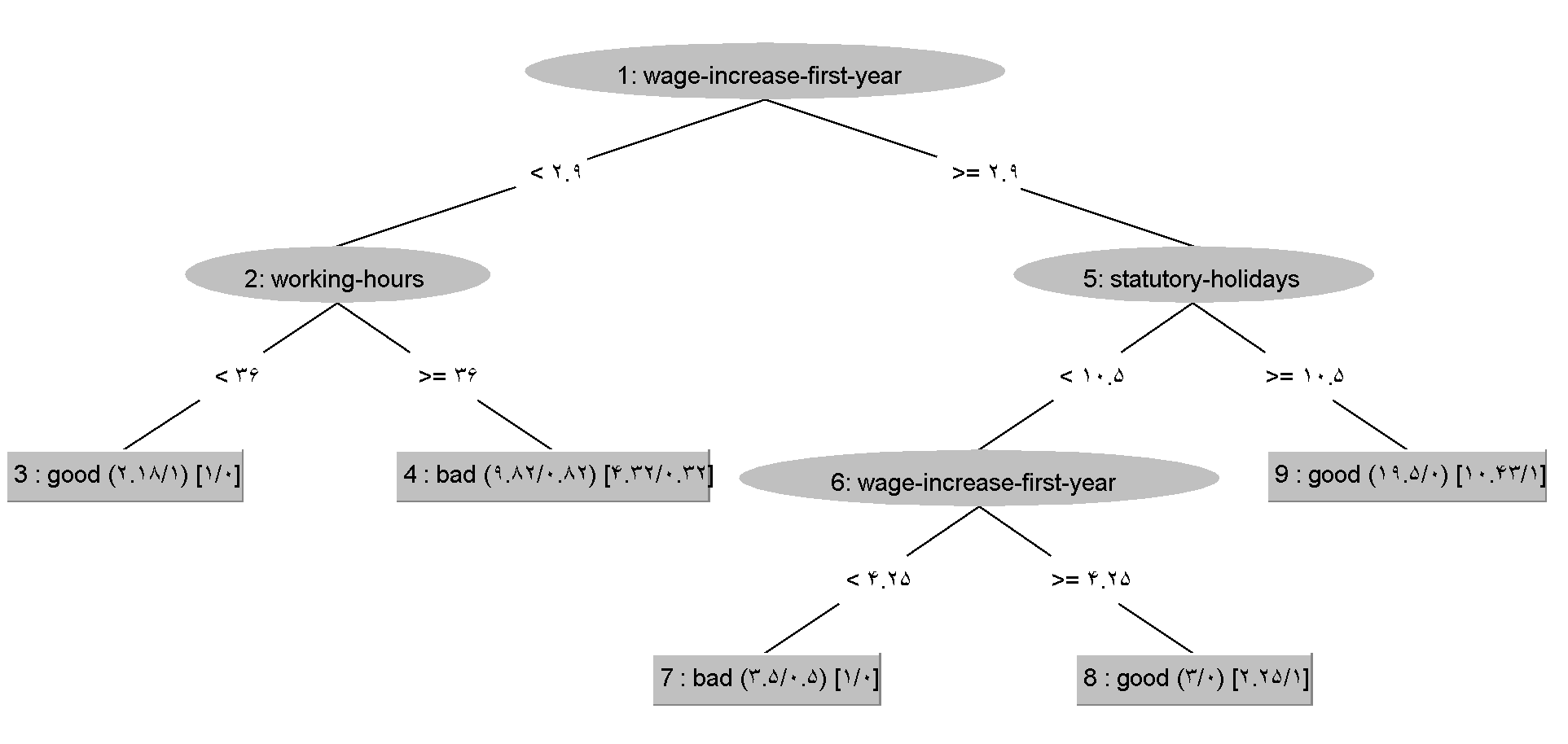


درخت نهایی برابر خواهد بود با:

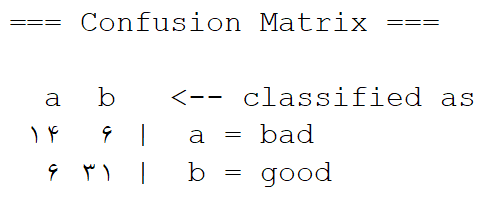


سوال ۵

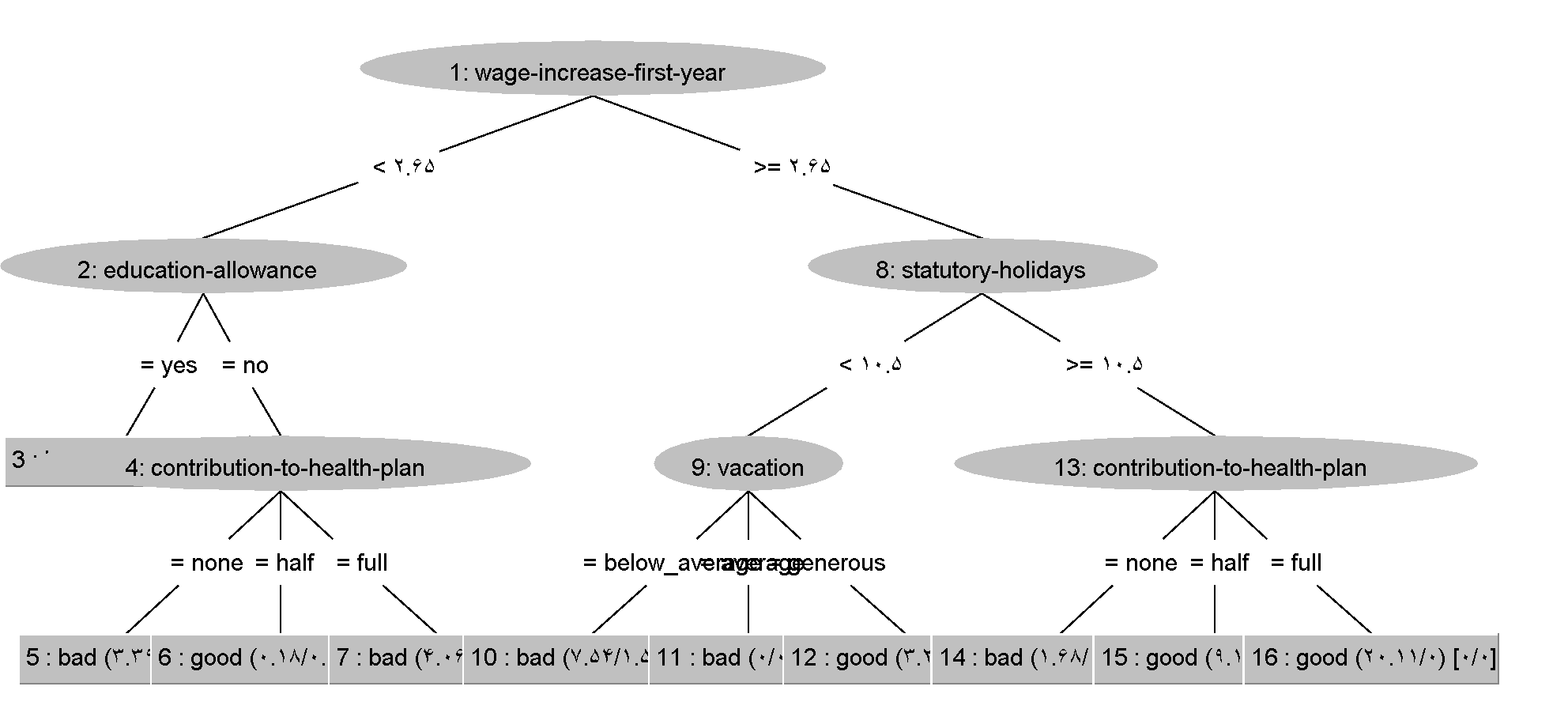
الف) برای این سوال از الگوریتم REPTree موجود در نرم‌افزار استفاده کردم و سایر پارامتر‌ها را تغییر ندادم.



ب)



ج)



د) مطابق انتظار درخت هرس‌نشده تعداد گره بیشتری دارد. همچنین می‌توان مشاهده کرد که درصد خلوص در گره‌های انتهایی درخت هرس‌نشده بیشتر است.

بخش دوم: پرسش‌های پیاده‌سازی

سوال ۱

الف)

ب) معیار‌های دقت برای بخش آموزشی و تست به شرح زیر است:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| F1 | دقت (Precision) | پوشش (Recall) | صحت (Accuracy) |  |
| 78.61٪ | 79.8٪ | 77.45٪ | 79.02٪ | مجموعه تست |
| 87.01٪ | 87.82٪ | 86.21٪ | 86.34٪ | مجموعه آموزشی |

ج) معیار‌های دقت برای بخش آموزشی و تست با حذف ویژگی chol به شرح زیر است:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| F1 | دقت (Precision) | پوشش (Recall) | صحت (Accuracy) |  |
| 79.21٪ | 80.81٪ | 77.67٪ | 79.51٪ | مجموعه تست |
| 87.27٪ | 88.29٪ | 86.27٪ | 86.59٪ | مجموعه آموزشی |

معیار‌های دقت برای بخش آموزشی و تست با حذف ویژگی oldpeak به شرح زیر است:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| F1 | دقت (Precision) | پوشش (Recall) | صحت (Accuracy) |  |
| 79٪ | 79.8٪ | 78.22٪ | 79.51٪ | مجموعه تست |
| 85.88٪ | 85.48٪ | 86.29٪ | 85.37٪ | مجموعه آموزشی |

با بررسی جداول این قسمت به نظر می‌رسد که حذف هر دو ویژگی می‌تواند دقت تست را بهبود دهد. از بین این دو حذف chol بهبود بهتری را به وجود می‌آورد.

سوال ۲

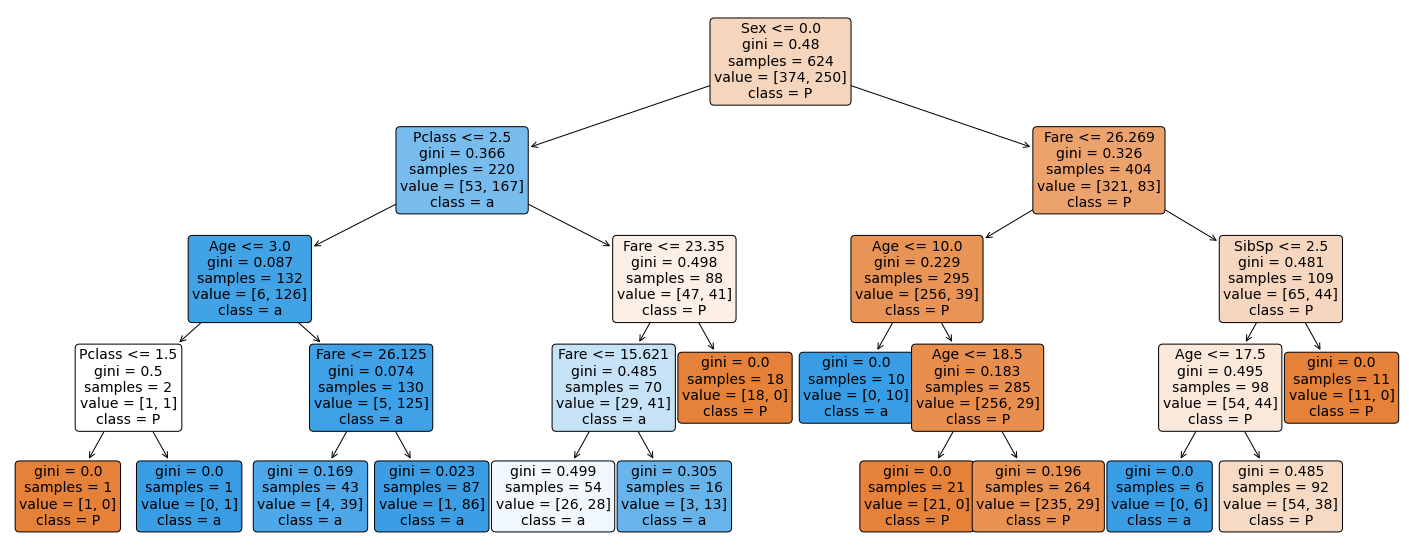
برای پیاده‌سازی این سوال، از پیاده‌سازی سابق خودم در دوره کارشناسی استفاده مجدد کردم.[[3]](#footnote-3)

در عین حال توجه کنید که مطابق اعلام تدریس‌یاران برای پیاده‌سازی این سوال از کتابخانه آمده برای درخت تصمیم استفاده کردم.

نهایتا توجه کنید که برای تعیین پارامترهای مدل مجموعه آموزشی را به دو مجموعه آموزشی و validation شکستم. ( 70٪ برای train و ۳۰٪ برای validation)

الف) ابتدا قبل از تصمیم برای مدیریت مقادیر گم‌شده، باید ببینیم کدام ویژگی‌ها باید حذف شود. در این مجموعه‌داده ویژگی‌های PassengerID،‌ Name‌ و Ticket‌ به نوعی نقش ID‌ دارند و باید حذف شوند. همچنین ویژگی Cabin مقادیر گم‌شده زیادی دارد و در عین حال چندان مفید نیست. برای سایر ویژگی‌ها عددی مانند Age‌ از میانگین همان ویژگی برای سایر داده‌ها استفاده کردیم. برای ویژگی Embarked هم تنها دو داده بدون مقدار بودند که برای آن‌ها بدون دلیل یک مقدار تصادفی درنظر گرفتیم.

ب) درخت تصمیم (نسبتا) بهینه را می‌توانید در تصویر زیر مشاهده کنید. عمق درخت ۴ و معیار خلوص Gini است. پس از آموزش داده‌های آموزش و Validation‌ را پیش‌بینی کردیم و در فایل train\_predicts.csv‌ ذخیره کردیم.

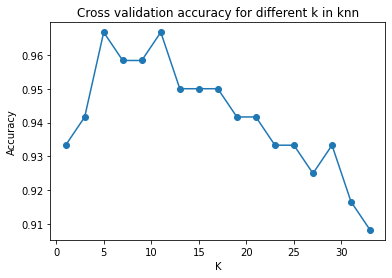


ج) مجموعه تست پردازش‌شده که شامل مقادیر پیش‌بینی شده هستند در فایل test\_predicts.csv آورده شده است.

سوال ۳

ابتدا باید اشاره کنم برای این مسئله ابتدا داده‌ها را shuffle‌ کردم و سپس 80٪ داده‌ها را برای آموزشی و 20٪ آن را برای تست در نظر گرفتم. توجه کنید که الگوریتم Cross Validation‌ روی قسمت‌های مختلف مجموعه آموزشی اعمال شده است.

الف) در نمودار زیر میزان صحت به ازای k های مختلف آورده شده است. به نظر می‌رسد مقدار k برابر با 5 یا 11 بهترین نتیجه را داشته است. برای قسمت بعد مقدار k را برابر با ۱۱ درنظر می‌گیرم. در اینجا لازم است این نکته را متذکر شوم که با بررسی‌ها انجام شده و shuffle کردن‌های متنوع امکان تغییر k پیشنهادی وجود دارد!



ب) برای مجموعه آموزشی میزان صحت برابر با 96.67٪ است و برای مجموعه تست میزان صحت برابر با 93.33٪ است. در دو ماتریس درهم‌ریختگی زیر می‌توانید ارزیابی دقیق‌تری داشته باشید.

ماتریس درهم‌ریختگی مجموعه آموزشی

|  |  |  |  |
| --- | --- | --- | --- |
| Iris-virginica | Iris-versicolor | Iris-setosa |  |
| 0 | 0 | 39 | Iris-setosa |
| 1 | 41 | 0 | Iris-versicolor |
| 36 | 3 | 0 | Iris-virginica |

ماتریس درهم‌ریختگی مجموعه تست

|  |  |  |  |
| --- | --- | --- | --- |
| Iris-virginica | Iris-versicolor | Iris-setosa |  |
| 0 | 0 | ۱۱ | Iris-setosa |
| 1 | ۷ | 0 | Iris-versicolor |
| ۱۰ | ۱ | 0 | Iris-virginica |

1. https://stats.stackexchange.com/questions/105979/is-knn-a-discriminative-learning-algorithm [↑](#footnote-ref-1)
2. https://stats.stackexchange.com/questions/12421/generative-vs-discriminative [↑](#footnote-ref-2)
3. https://gitlab.com/aut-data-mining/titanic/ [↑](#footnote-ref-3)