تمرین دوم پیاده‌سازی داده‌کاوی

استاد مزلقانی

عطیه براتی نیا ۹۶۳۱۰۱۰

سوال ۶

مرحله‌ی اول: آماده‌سازی داده‌ها

ویژگی‌های Name, PassengerId, Ticket برای هر فرد یکتا است و نمیتوان از روی این ویژگی‌ها نمیتوان افراد را دسته‌بندی کرد و تشخیص داد که آیا فرد زنده میماند یا نه، در نتیجه این ویژگی‌ها در درخت استفاده نمیشوند. ویژگی Cabin در اکثر موارد خالی است بنابراین از این ویژگی نیز نمیتوان تصمیم درستی گرفت.

ویژگی‌ها باید عددی باشند به همین دلیل male => 0 و female => 1 از مقادیر ستون sex تغییر داده میشوند. از ستون Embarked نیز C=>1, Q =>2, S=>3 تبدیل میشوند.

درخت تصمیم نباید ویژگی خالی داشته باشد. بعضی از مقادیر ستون Age خالی است که آنها را با میانگین سن افراد پر میکنیم. Fare نیز در برخی موارد خالی است که آن را نیز با میانگین بقیه موارد آن ستون پر میکنیم. برخی موارد در ستون Embarked نیز خالی است، در این ستون نمیتوان میانگین گرفت به همین دلیل موارد خالی را با mode این ستون پر میکنیم.

مرحله دوم: آموزش و دسته‌بندی

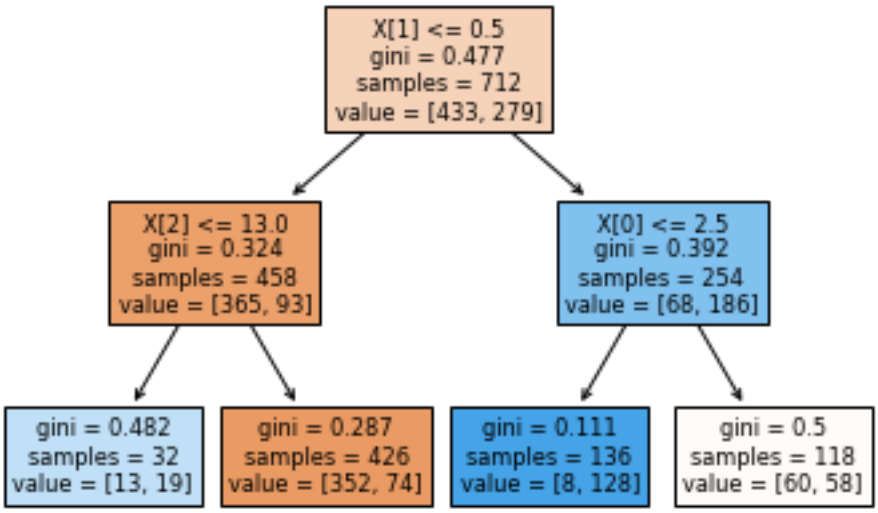
از متد features_train, features_test, label_train, label_test = train_test_split(df, label, test_size = 0.2) برای تقسیم‌بندی به دو دسته تست و آموزش استفاده شد.

*توجه: از آنجایی که این متد به طور رندم مقدارهای تست و آموزش را انتخاب میکند هر اجرای برنامه دقت متفاوتی میدهد، به طور مثال در یک دور تست برای عمق ۶ دقت ۰.۸۳ نیز به دست آمده بود. مقادیری که در ادامه آمده‌اند بر روی یک نمونه تست و آموزش ثابت اندازه‌گیری شده است.

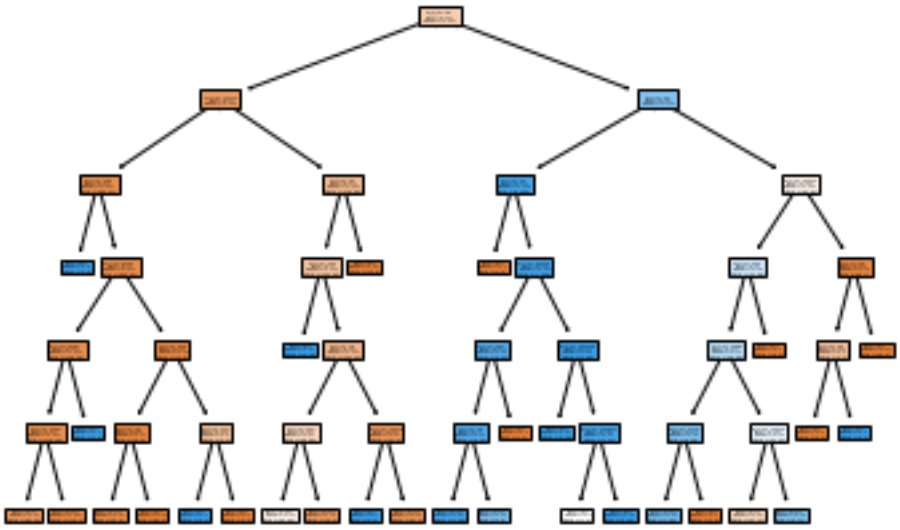
در جدول زیر حالت‌های مختلف مربوط به تابع تقسیم gini به همراه درصد دقت آورده شده‌است.

ردیف	عمق درخت	تابع تقسیم	دقت
1	2	gini	0.743
2	3	gini	0.782
3	6	gini	0.793
4	7	gini	0.787
5	20	gini	0.759
6	نامحدود	gini	0.765

دو نمودار مربوط به رسم درخت به عمق ۲ و عمق ۶ برای نمونه در ادامه آورده شده است.



شکل ۱ رسم درخت با تابع gini با حداکثر عمق ۲

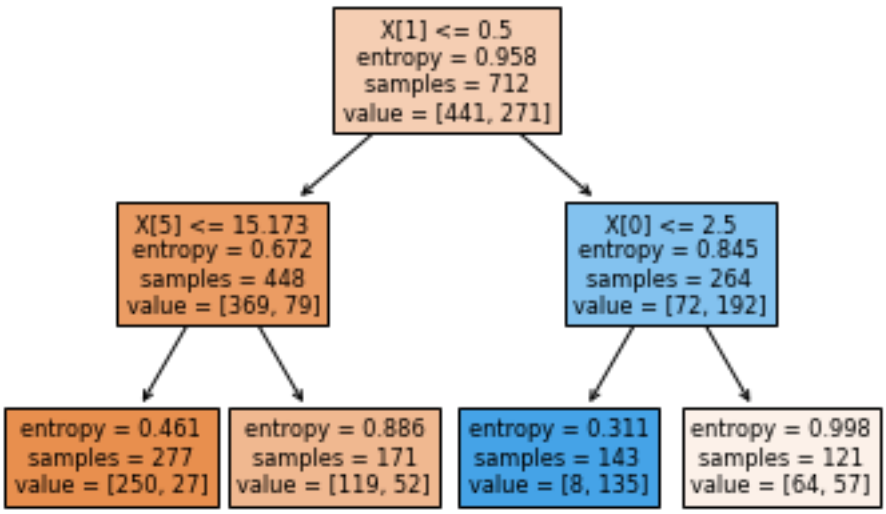


شکل ۲ رسم درخت با تابع gini با حداکثر عمق ۶

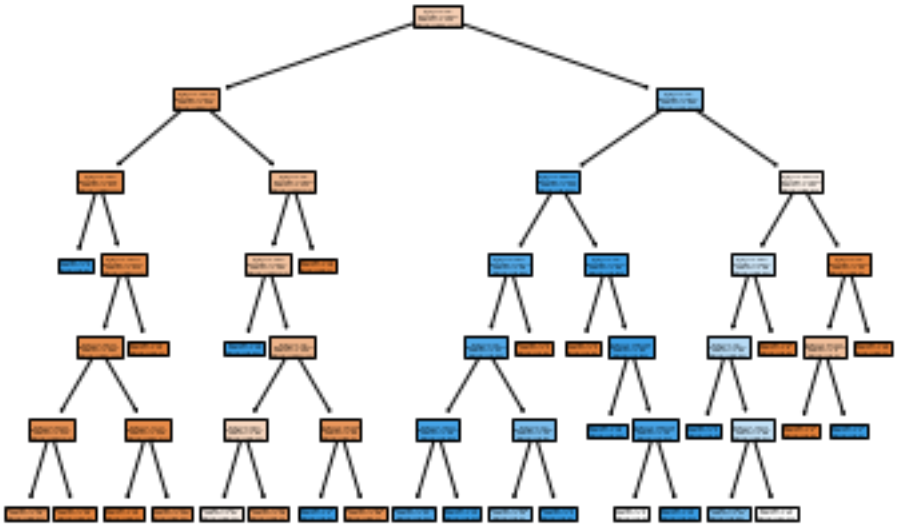
در جدول زیر حالت‌های مختلف مربوط به تابع تقسیم entropy به همراه درصد دقت آورده شده‌است.

ردیف	عمق درخت	تابع تقسیم	دقت
1	2	entropy	0.743
2	3	entropy	0.787
3	6	entropy	0.793
4	7	entropy	0.770
5	20	entropy	0.720
6	نامحدود	entropy	0.737

دو نمودار مربوط به رسم درخت به عمق ۲ و عمق ۶ برای نمونه در ادامه آورده شده است.



شکل ۳ رسم درخت با تابع entropy با حداکثر عمق ۲



شکل ۴ رسم درخت با تابع entropy با حداکثر عمق ۶

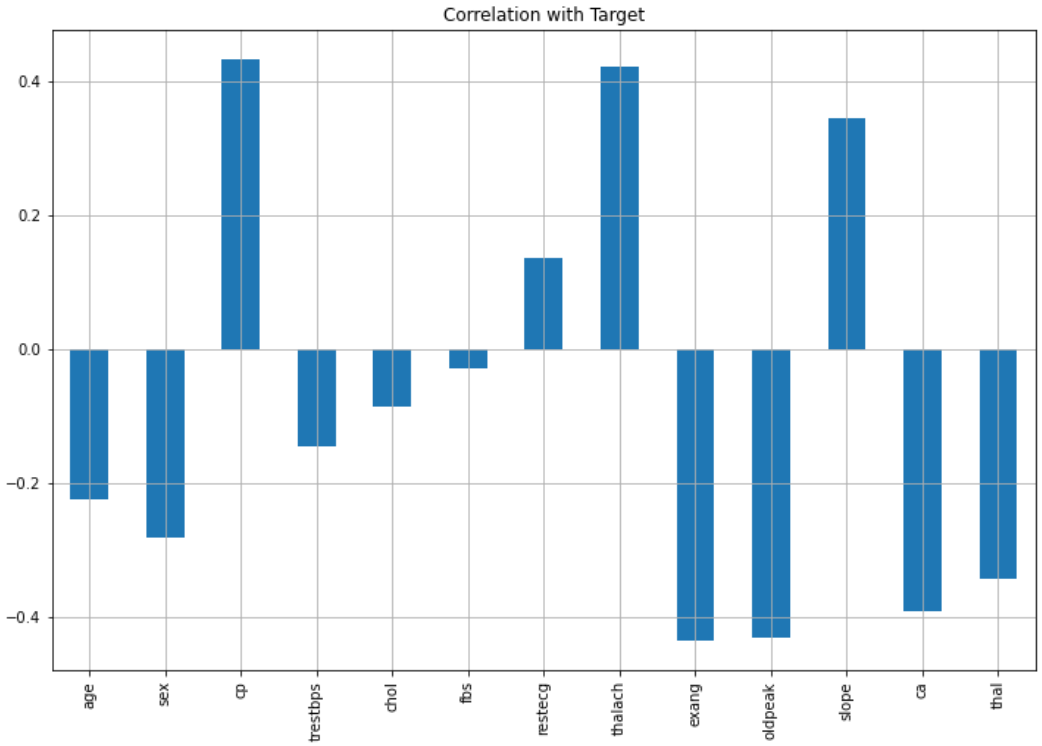
سوال ۷

مرحله‌ی اول: تحلیل و آماده‌سازی داده‌ها

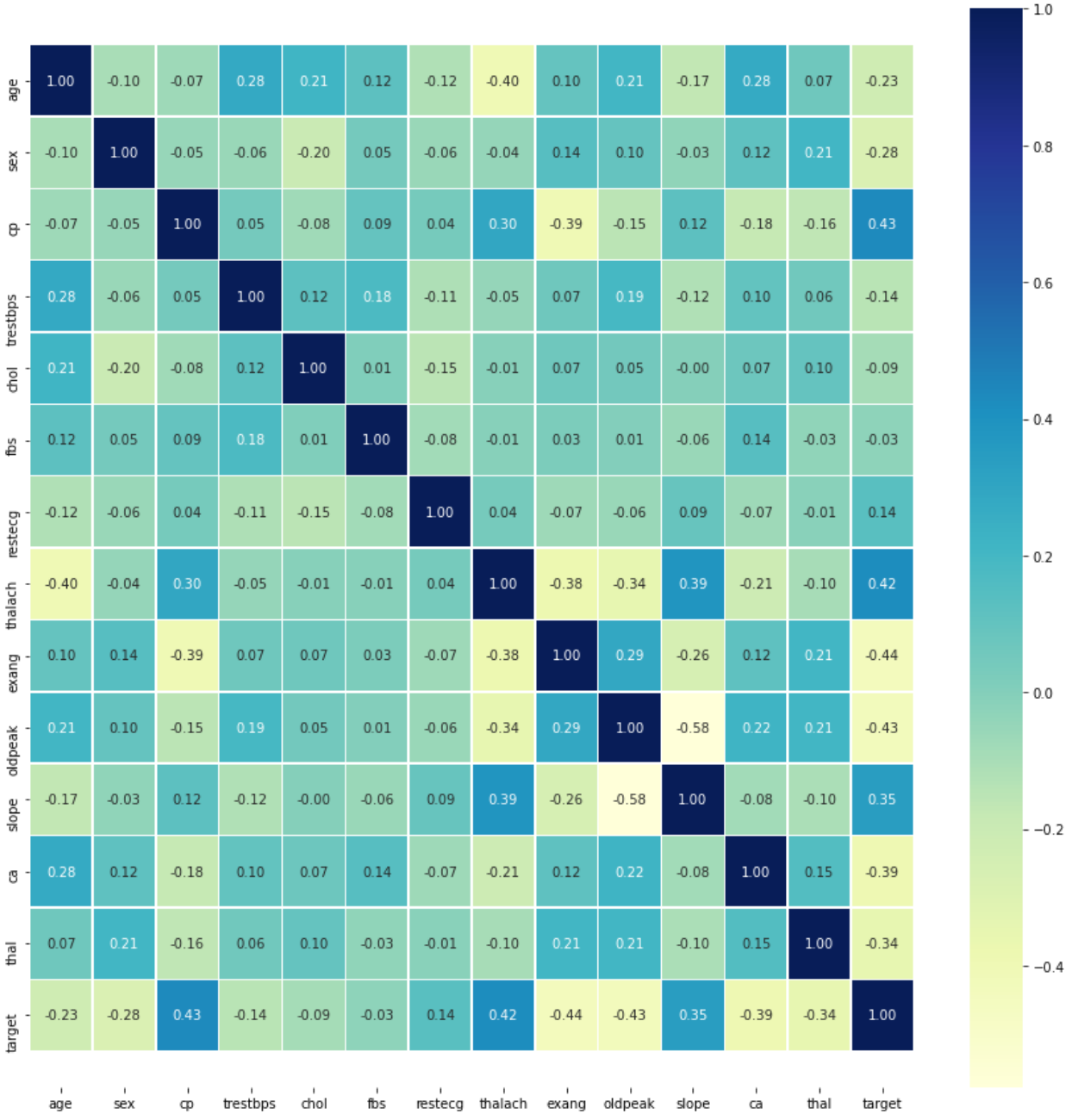
در ابتدا داده‌ها را به دو قسمت آموزشی و تست تقسیم میکنیم.

سپس ستون‌ها را از لحاظ مقادیر دسته‌بندی میکنیم که با مقادیر هر ستون بهتر آشنا شویم. سپس بین ویژگی‌ها با یکدیگر کورلیشن میگیریم تا ببینیم ارتباط خطی دارند یا نه. این ارتباط را به صورت دو نمودار یکی کورلیشن بین ویژگی ها با یکدیگر و دیگری کورلیشن بین هر ویژگی با target نشان میدهیم. به نظر میرسد fbs کورلیشن پایینی دارد ولی باید این مطلب را در نظر گرفت که کورلیشن فقط ارتباط خطی را بررسی میکند و ممکن است این ویژگی ارتباط غیرخطی با target داشته باشد (این مطلب از روی تست‌های متعددی که گرفته شد مشخص شد)، به همین دلیل ویژگی‌ای را حذف نمیکنیم.

در آخر به دلیل محدوده متفاوتی که مقادیر با یکدیگر دارند ویژگی‌ها را استاندارد میکنیم.



correlation with target شکل ۵



مرحله‌ی دوم: ساخت و آموزش مدل

داده‌ها را با KNN با k=6 ، فاصله منهتن و وزن‌دهی بیشتر به همسایه‌های نزدیک‌تر، مدل میکنیم. احتمال بالای ۰.۸ است که در یک نمونه مقدار آن ۰.۸۱۹ اندازه‌گیری شد.

همچنین داده‌ها را با naïve bayes مدل میکنیم که احتمال آن به طور میانگین ۰.۸ است و بر روی نمونه‌ی یکسان که مدل قبل ۰.۸۱۹ شد، این مدل نیز ۰.۸۱۹ شد.(مقدار این دو مدل همیشه برابر نیست و این بار استثنائا برابر شدند.)