سوال1)

در ابتدا چک می کنیم که دیتای nan و جود دارد یا نه بعد ستون های مدنظر تفکیک کرده

بعد میام تابع forward که یه threshold\_in براش در نظر میگریم اگر کوچک تر از p-value اونو بهترین ستون مدنظر می گیریم

2)

بعد مابقی ستون ها را حذف کرده و بهترین ستون هارو train و test می کنیم

همه ستون های train و test رو در تابع standardScaler تو یه رنج برده

بعد train و test در lntglogistic Regression که روش classification هست فیت کرده

بعد همین کارا رو به کمک تابع predict بر روی test مون هم انجام میدبم

بعد میایم به کمک تابع concatenate مقداری که پیش بینی شده و مقدارtest مون رو کنارهم قرار میدیم

خروجی confusion\_matrix ما یه ماتریسی هست که جمع درایه های اصلی آن و جمع همه اعداد ماتریس score مون هست

3) سوال 3 فکر کنم منظور بد فهمیدم اشتباه نوشتم

4) مثل دستور سوال دو پیش میریم

6)

الف) در روش bining میام کمترین مقدار و بیشترین مقدار ستون رو گرفته و به کمک linespace 4 عدد رندوم به طور مساوی گرفته و ستون جدید میسازیم

ب) ستون هایی که بیش از 2 حرفی هستن (مثل pracing\_range) به کمک oneHotencoder به صورت ماتریس ویا به کمک تابعget\_dummies ستون به ستون عددی می کنیم

8) بوت استرپینگ میاد با تغییر رندوم سطر یا حذف یه سطر و جایگزینی آن جدوال مختلفی ساخته و و مقدارواریانس هرکدوم مشخض کرده و مجموع آن ها را تقسیم بر میانگین کرده تا حواب واقعی نزدیک تر شیم

در cross میایم دیتارو میشکافیم و یه قسمت به عنوان train و مابقی test و دوباره میایم قسمت دیگر train و مابقی تست می کنیم

بوت استرپینگ برای مواقعی که دیتای کمی داریم مناسب هست

9)

داده ها به طور تصادفی به دو بلوک تقسیم میشوند و 5 بار ارزیابی می کنیم

10)

درخت ID3 :

این یه الگوریتم ساده به عنوان معیار تفکیک هست و هیچ فرایند هرس کردن بکار نمیگیرد و مقادیراسمی و مفقوده رو مورد توجه قرار نمی دهد

درخت C4.5:

این الگوریت توسعه یافته ID3 است و GAIN RATIO به عنوان معیار تفکیک در نظر گرفته می شود . در عمل این الگوریتم مشخصه اسمی را مد نظر گرفته می شود

درخت CART :

برای برقراری درخت های رگرسیون و دسته بندی از ای الگوریتم استفاده می شود . این الگوریتم توانایی تولید درخت های رگرسیون را دارد . این الگوریتم برای تفکیک کننده میزان مینیمم مربع خطاا جست و جو می کند

13)

هدف ما از Pruning هرس کردن حذف جواب های غیر ممکن است . زمانی که یه درخت تصمیم ساخته می شود تعدادی از شاخه های بی خود در داده ای اموزشی ساخته می مشود که ناشی از داده های پرت و یا نویز است