گزارش پروژه درخت تصمیم
نرگس سیفی 99442173

پروژه من شامل دو فایل نوتبوک است که در اولی داده ها را پیش پردازش کردم و خروجی را در فایل csv ذخیره و در فایل بعدی استفاده کردم.

توضیحات لازم برای اجرا: فقط همه ی فایل ها باید در یک پوشه باشند.

در فایل decisionTree، دو کلاس Node و DecisionTree نوشته شده که پیاده سازی الگوریتم درخت تصمیم در آنها انجام شده. (برای شاخص های انتروپی و gini index هم توابع جدا نوشته ام)

بعد از پیاده سازی، با یک مثال کوچک (داده ی آب و هوا) عملکرد درخت را تست کرده ام و با کمک این مثال باگ های سینتکسی و الگوریتمی را حل کردم.

در مرحله بعدی داده هایی که از قبل پیش پردازش شده اند را خواندم و به دو قسمت test , train تقسیم کردم. سپس دو مدل از کلاس DecisionTreeخود ساختم که در یکی از انتروپی و در بعدی از Gini برای ساخت درخت استفاده کرده ام. (با استفاده از پارامتر ورودی gainState)

و سپس precision , accuracy را برای هر دو این مدل ها محاسبه کردم. طبق این نتایج، برای بدست آوردن درخت بهتر، باید threshold را در درختی که از gini استفاده میکند کمتر از درختی که انتروپی استفاده میکند در نظر میگرفتم. (با مقدار 0.05 که انتروپی به خوبی جواب میداد، جینی درختی با تنها یک راس میساخت.)

پیش پردازش:

در این فایل، ابتدا missing value ها را بررسی کردم و بعضی از آنها را پر کردم و بعضی را چون خیلی تعدادشان کم بود حذف کردم. در ادامه با کشیدن نمودار های مختلف و بررسی داده ها، ستون های عددی را به ستونهای کتگوری تبدیل کردم و ستونهای کتگوری که تعداد دسته هایشان خیلی زیاد بود را به دسته های کمتری تبدیل کردم تا درخت بازدهی بیشتری داشته باشد. همچنین لیبل y را به 0و 1 تبدیل کردم با استفاده از labelEncoder که بعدا بتوانم از توابع precision و ... استفاده کنم.

از جمله چالش هایی که بر خوردم این بود که بعضی مقادیر در داده های train نبود اما در داده های test دیده میشد و درخت من نمیتوانست آن سطر از داده را پیش بینی کند، با سرچ در اینترنت و کمک از gpt، به این نتیجه رسیدم که برای هر راسی مقدار اکثریت برچسب داده هایی که با آنها train شده را نگه دارم (فارغ از اینکه راس برگ است یا خیر) و در جایی که به این شرایط ویژه برخوردم از این اکثریت برچسب استفاده کنم برای آن سطر داده.

در داده هایی که ما داشتیم این کار اثر منفی بزرگی روی خروجی نمیگذاشت چون از بین 11هزار و .. داده ی تست، تنها 11 تا از سطر ها چنین مشکلی داشتند.