فاطمه سادات باقری – 400522355

پروژه درخت تصمیم

در ابتدا به توضیحات مربوط به معیار های Entropy و Gini Index می پردازیم.

یک درخت از گره‌ها تشکیل شده است، و این گره‌ها، با جستجو برای تقسیم بهینه ویژگی‌ها انتخاب می‌شوند. برای این منظور، معیارهای مختلف وجود دارد. این پارامتر عملکردی است که برای اندازه‌گیری کیفیت یک تقسیم در هر گره استفاده می‌شود و به کاربران امکان انتخاب بین Gini یا Entropy را می‌دهد.

- محاسبه Gini impurity :

$$GiniIndex = 1 - \sum_j p_j^2$$

- محاسبه Entropy :

$$Entropy = -\sum_j p_j \cdot log_2 \cdot p_j$$

- مقایسه Gini و Entropy :

مقایسه بین معیار جینی و آنتروپی برای تقسیم داده‌ها در درخت تصمیم به شرح زیر است:

1. Gini index:

- معیار جینی یک معیار برای اندازه‌گیری خلوص یک تقسیم است. زمانی که ایندکس جینی صفر است، به این معناست که تمام نمونه‌های یک گروه به یک دسته تعلق دارند. به عبارت دیگر، ایندکس جینی یک اندازه‌گیری از این است که چقدر داده‌ها در هر گروه یکسان هستند.
- استفاده از اندیس جینی عموماً برای مسائلی که توزیع داده‌ها مهم است و اهمیت آنها برابر است مناسب است، مانند درخت‌های تصمیم برای دسته‌بندی.

2. Entropy:

- آنتروپی نشان‌دهنده میزان ناهمگنی یا تنوع داده‌ها است. زمانی که آنتروپی صفر است، به این معناست که همه داده‌ها به یک کلاس یا دسته تعلق دارند. به عبارت دیگر، آنتروپی کمتر به معنای تمیزتر بودن و داده‌های مشابه در هر گروه است.

- استفاده از آنتروپی معمولاً برای مسائلی که تفاوت بین داده‌ها مهم است و اهمیت آنها متفاوت است مناسب است، مانند درخت‌های تصمیم برای تقسیم‌بندی که تمایل به تعادل بین گروه‌ها دارند.

بنابراین، انتخاب بین معیار جینی و آنتروپی بستگی به مسئله مورد نظر و ویژگی‌های داده دارد. در بعضی موارد، هر دو معیار می‌توانند به خوبی عمل کنند، در حالی که در موارد دیگر، یکی از آنها ممکن است بهترین انتخاب باشد بر اساس خصوصیات داده و هدف مسئله.

- در درخت تصمیم نهایی ساخته شده نتایج حاصل از این دو تقریبا نزدیک هم ومعمولا حدود **97٪** می باشد.

- در درخت تصمیم ابتدایی که درخت تصمیم را به صورت Binary ایجاد میشد با استفاده از رندوم انتخاب کردن feature ها و همچنین با قرار دادن محدودیت برای عمق درخت، همچنین مینیمم قرار دادن تعداد feature ها برای انتخاب، قصد داشتم که از بزرگ شدن درخت جلوگیری کنم و تقریبا هم دقت خوبی به دست آوردم که با 30 بار ران گرفتن میانگین آن عددی در حدود 80٪ بود.

حال به توضیح روند پروژه می پردازیم:

1. در ابتدا ویدیوی های موجود در youtube به توضیح gini index و entropy پرداخته بودند را مشاهده کردم.
2. سپس چندین ویدیو که چگونگی پیاده سازی درخت تصمیم را توضیح داده بودند دیدم و بعد دیدن این ویدیو ها، شروع به زدن پروژه کردم.
3. در ابتدا داده ها را شروع به تحلیل کردم و ستون هایی که در درخت تصمیم نقش اساسی برای انتخاب هایمان نداشتند را حذف کردم.
4. در ابتدا این پروژه را را که درخت همیشه binary باشد پیاده سازی کردم و مشکلی که در این درخت داشتم accuracy آن چندان دقیق نبود که مجبور به حذف محدودیت هایی که برای درخت قرار داده بودم.
5. پس از این که در گروه گفته شد که درخت ما نباید binary باشد و باید گسسته سازی حتما صورت گیرد، در ابتدا به گسسته سازی داده ها پرداختم و سپس این درخت را به صورت none binary ساختم.
6. در نهایت به visualization درخت پرداختم که به دلیل این که تعداد فرزندان هر نود زیاد میشد، شکل به طور مناسبی نمایش داده نمی شد، و مجبور به تست چند حالت شدم که نتایج نهایی در pdf ذخیره شده اند. نمونه ها در فایل های جداگانه قرار داده شده است.