ارمین رحمتی

993112050

پروژه انتقال داده

دکتر کاظمی تبار

همان طور که مشخص است کد رو دو بار با انتخاب 2000و 5000 نمونه تصادفی رندوم کد را اجرا میکنیم.

استفاده ازrandom_state برای مقایسه عملکرد آزمایش با اندازه ردیف های مختلف در حالی که عوامل دیگر را ثابت نگه داشته ایم، مفید است در واقع فرآیندهای تصادفی انتخاب نمونه را قابل تکرار می کنیم.

در تابع قسمتpreprocess_data داده ها را برای آموزش مدل یادگیری ماشین آماده میکنیم در و اقع ستون و سطر هایی که درمحاسبات تاثیر ندارند را حذف و همین طور در مورد ستون ها با داده های خالی میانه ان ستوت رو با ان ها جایگزاری البته ستون هایی که شامل مقدار عددی نیستند از مد استفاده کردم و همین طور عمل نرمال سازی را برای مقادیر عملی انجام دادم.

شایان ذکر است این دو عمل در دقا مدل که با درخت تصمیم ساخته شده تاثیری ندارد و در مسایلc nnکه با فاصله ها سر کار داریم مفید است.

سپس تقسیم مجموعه داده به مجموعه های آموزشی و آزمایشی را داریم (train_test_split).

حال به ساخت درخت تصمیم با عمق دلخواه 5 میپردازیم.

سپس در find_best_split ویژگی را برمیگردانیم که حداکثر بهره اطلاعات را ارائه میکند، که بهترین ویژگی برای تقسیم کردن است در واقع root درخت تصمیم را تشکیل میدهد.

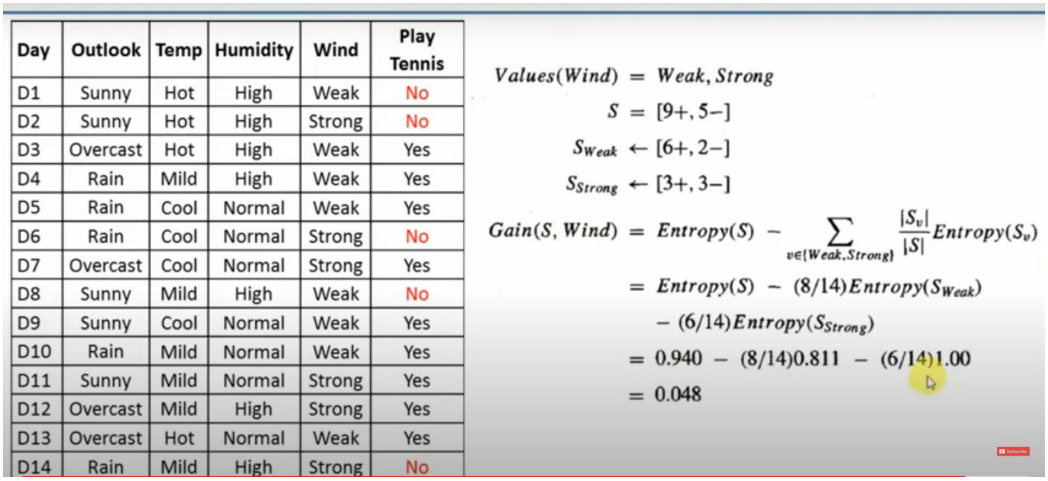
ویژگی که کمترین ناخالصی یا ابهام را دارد به عنوان مهم ترین ویژگی انتخاب میشود.

حداکثر بهره اطلاعات بر اساس سود اطلاعات که تفاوت بین آنتروپی قبل از تقسیم و مجموع وزنی آنتروپی ها پس از تقسیم است بدست می اید.

$$E=-\sum pi*log2pi$$

$$Gani(T,X) = Entropy(T) - Entropy(T,X)$$

یک نمونه از محاسبات را در شکل زیر برا ویژگی wind مشاهده میکنیم:

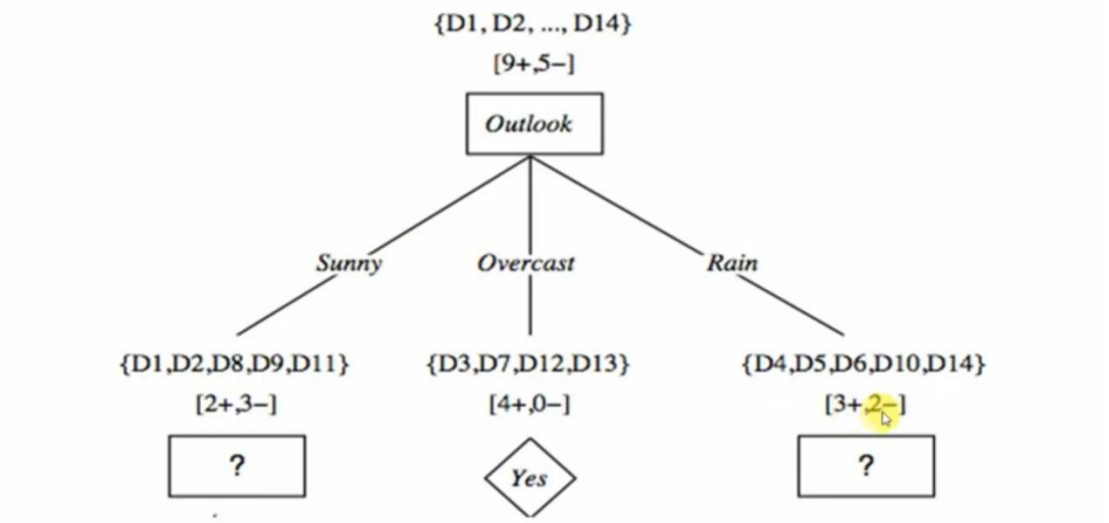


Day	Outlook	Temp	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$Values(Wind) = Weak, Strong$$
$$S = [9+, 5-]$$
$$S_{Weak} \leftarrow [6+, 2-]$$
$$S_{Strong} \leftarrow [3+, 3-]$$
$$Gain(S, Wind) = Entropy(S) - \sum_{v \in \{Weak, Strong\}} \frac{|S_v|}{|S|} Entropy(S_v)$$
$$= Entropy(S) - (8/14) Entropy(S_{Weak}) - (6/14) Entropy(S_{Strong})$$
$$= 0.940 - (8/14)0.811 - (6/14)1.00$$
$$= 0.048$$

به ایجاد گره های فرزند در درخت تصمیم برای هر مقدار منحصر به فرد بهترین ویژگی میپردازیم در واقع مقادیر مختلف که در ستون مربوط به بهترین feature هستند تعداد شاخه هایی که از root ایجاد میشوند را مشخص میکنند.

برای مثال عکس زیر را در نظر داریم:



ویژگی outlook که دارای 3 زیرشاخه بوده که مقدار ستون satisfication هر کدارم زیر ان مشخص شده همان طور که مشخص است در حالت overcast به yes رسیدیم یعنی فقط یک مقدار satisfication رو داشت پس این برگ است.

برای هر زیرشاخه، یک زیردرخت با استفاده از نقاط داده ای که دارای آن مقدار ویژگی خاص هستند ساخته می شود که این عمل محاسبات بهترین ویژگی تنها برای این سطر ها به صورت بازگشتی ادامه میابد تا زمانی که یک شرط توقف برقرار شود، مانند رسیدن به حداکثر عمق مشخص یا داشتن برگ.

ساختار حاصل یک درخت تصمیم است که در آن هر گره نشان دهنده یک ویژگی، هر شاخه نشان دهنده یک مقدار ویژگی، و هر برگ نشان دهنده یک برچسب کلاس پیش بینی شده است.

دقت مدل به عنوان اولین خروجی ان برچسب ها را با استفاده از طبقه بندی کننده پیش بینی میکند ، دقت پیش بینی ها را محاسبه میکند.

Accuracy for 2000 rows: 0.6025

Accuracy for 5000 rows: 0.573

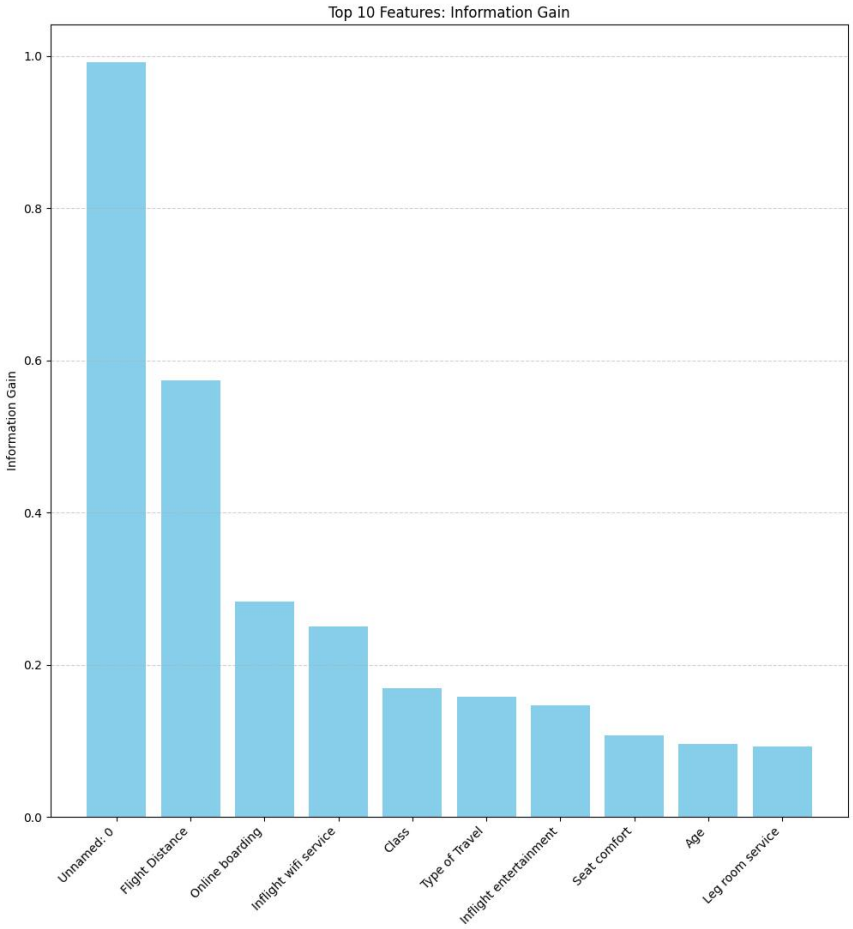
دلیل اینکه مقدار دقت برای تعداد سطر بیشتر کمتر شده است رندومنس در امنتخاب داده ها می باشد یعنی تعداد داده یکسانی از ابتدا تا 2000 و 5000 انتخاب نشده و کاملا رندوم است.

خروجی دیگر نمایش داده شده مربوط به بهره اطلاعات برای بهترین ویژگی در تابع calculate_and_save_information_gain بیاده سازی شده است که نتایج زیر را به همراه دارد.

2000 rows

results > rows_2000 > information_gain_results.csv

```
Feature,Information Gain
Unnamed: 0,0.9918412712305922
Flight Distance,0.5736638626874766
Online boarding,0.2834250393972697
Inflight wifi service,0.2499828026872648
Class,0.16933692895054864
Type of Travel,0.15780357742553208
Inflight entertainment,0.14671728584593335
Seat comfort,0.10685122675789827
Age,0.09531633100608727
Leg room service,0.09225640525234124
```

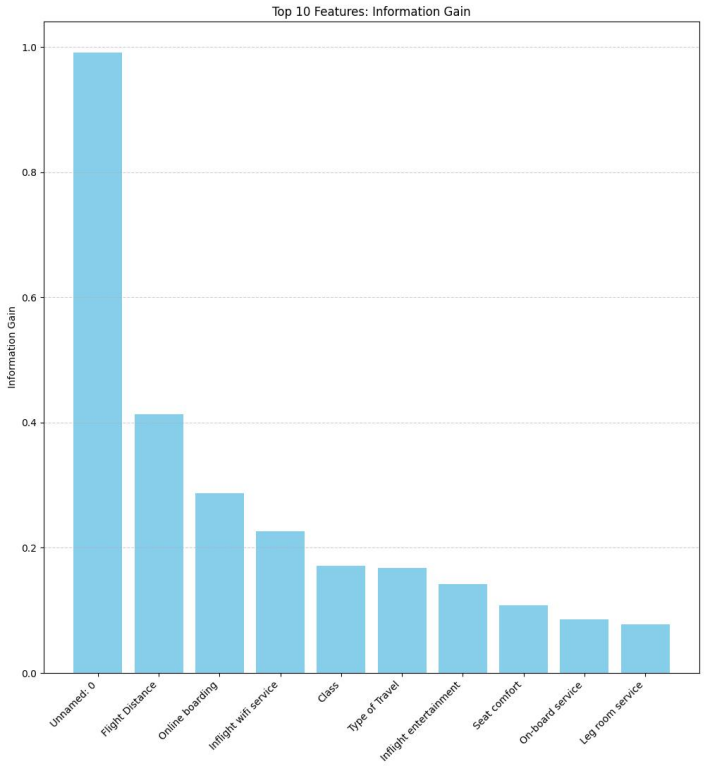


5000 rows

main.py | information_gain_results.csv | decision_tree.py

results > rows_5000 > information_gain_results.csv

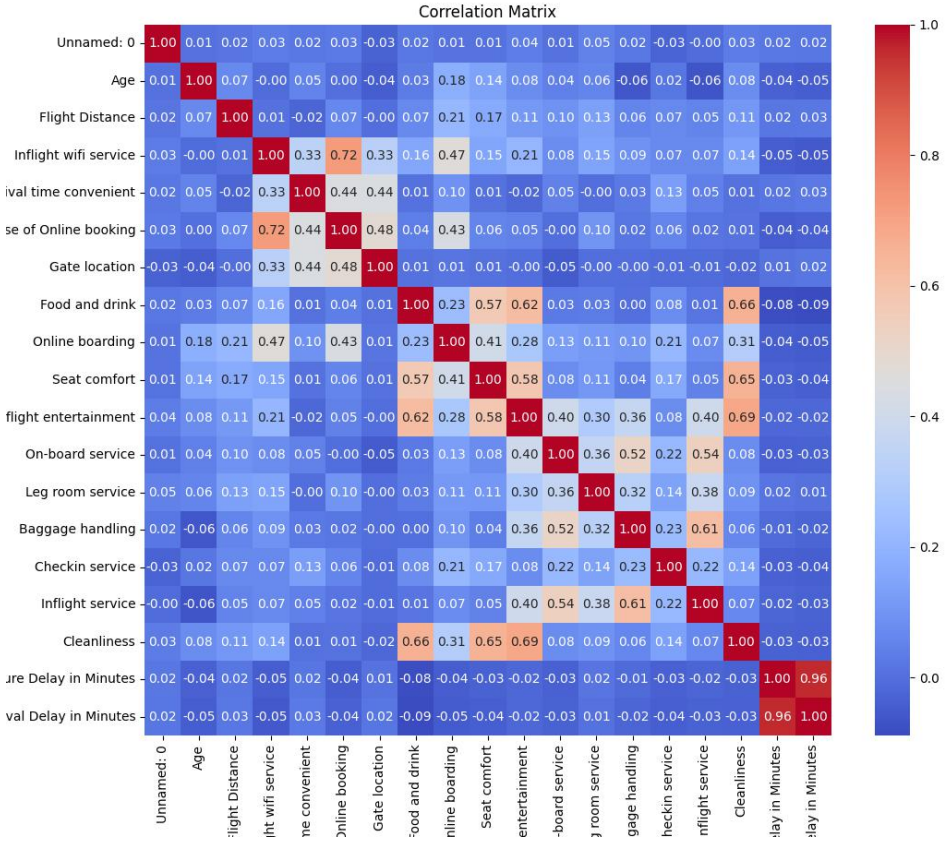
```
Feature,Information Gain
Unnamed: 0,0.9910133521048275
Flight Distance,0.413051009621751
Online boarding,0.2872070343751817
Inflight wifi service,0.22688733356513024
Class,0.1716318361581618
Type of Travel,0.16766944813565154
Inflight entertainment,0.1417353345872876
Seat comfort,0.10783217749197183
On-board service,0.08590124735211113
Leg room service,0.07762684144908716
```


Top 10 Features: Information Gain
Information Gain
1.0
0.8
0.6
0.4
0.2
0.0
Unnamed: 0
Flight Distance
Online boarding
Inflight wifi service
Class
Type of Travel
Inflight entertainment
Seat comfort
On-board service
Leg room service

خروجی دیگر ماتریس همبستگی را برای ستون های عددی در یک مجموعه داده محاسبه میکند و نقشه حرارتی حاصل را به عنوان یک تصویر ذخیره میکنیم.

این ماتریس برای تعیین کمیت و تجسم روابط خطی بین جفت متغیرهای عددی استفاده میشود .

یک همبستگی مثبت نشان می دهد که با افزایش یک متغیر، متغیر دیگر نیز تمایل به افزایش دارد، در حالی که همبستگی منفی نشان دهنده یک رابطه معکوس است.



خروجی دیگر نیز پراکندگی ستون ها یا همان فیچر ها میباشد که هرکدام جداگانه در پوشه distribution_plotsزخیره شده اند که نیاز به توضیح ندارد.

برای مثال پراکندی سن در دیتاست.