امیرحسین سلیمی

شماره دانشجویی : 400521432

پروژه یک

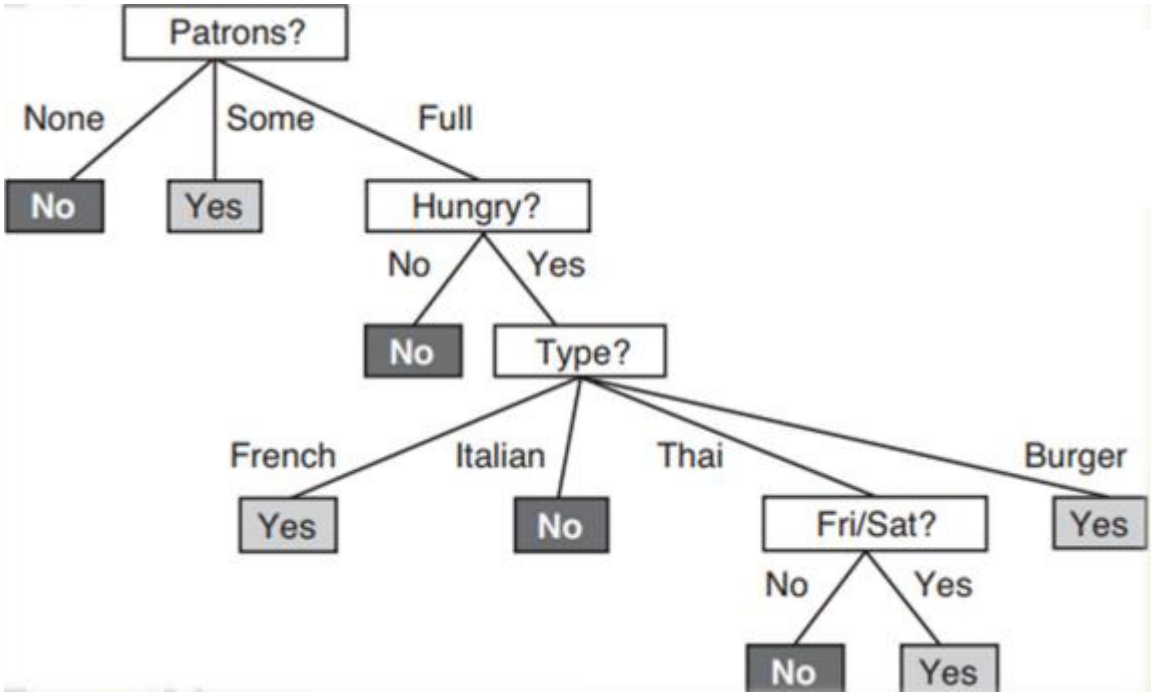
مرحله اول

درخت ساخته شده با استفاده از آنترپی :

در هر سطح نام ویژگی که برای جدا کردن استفاده میشود آمده است برای مثال در بالاترین سطح یا همان گره root بهترین ویژگی است که ممکن است انتخاب شود با توجه به information gain ویژگی pat بهترین ویژگی است.

در گام بعد اگر pat = none باشد کسانی که صبر نخواهند کرد فقط در این بخش خواهند بود و به عبارتی آنترپی آن صفر میشود و برای some هم همینطور با این تفاوت که در some فقط کسانی که صبر خواهند کرد قرار میگیرند.

اما در پارت Full از هر دو گروه داریم و بهترین ویژگی برای جدا کردن آنها نوع غذا خواهد بود که باعث میشود سه کلاس آنترپیشان صفر شود ولی در بخش Thai بازهم از هر دو نوع داریم که با توجه به ویژگی Fri آنها هم جدا میشوند. تصویر گرافیکی آن در صفحه بعد قابل مشاهده است.



درخت ساخته شده با Gini Index

```
Gini:
{
  "Pat": {
    "None": 0,
    "Some": 1,
    "Full": {
      "Hun": {
        "0": 0,
        "1": {
          "Type": {
            "Thai": {
              "Fri": {
                "0": 0,
                "1": 1
              }
            },
            "Burger": 1,
            "French": 0,
            "Italian": 0
          }
        }
      }
    }
  }
}
```

تفاوت Gini Index و Entropy

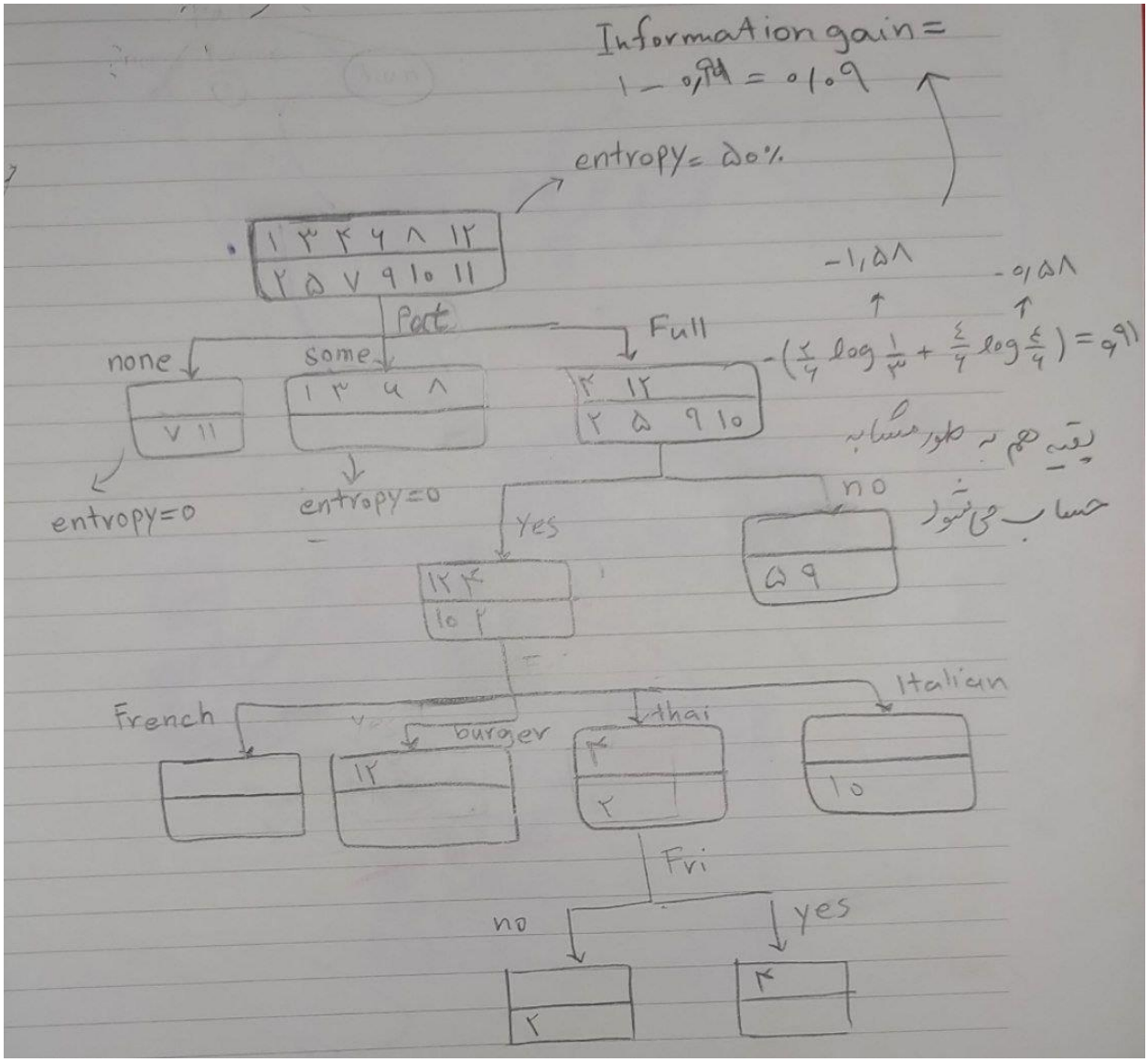
Gini index و entropy دو معیار برای اندازه گیری خلوص یا ناخالصی یک مجموعه داده هستند .این دو معیار اغلب در الگوریتم های درخت تصمیم برای انتخاب ویژگی مناسب برای تقسیم یک مجموعه داده استفاده می شوند.

Gini indexاحتمال اینکه یک نمونه به طور تصادفی انتخاب شده از یک مجموعه داده به طور نادرست طبقه بندی شود را اندازه گیری می کند .مقدار Gini index بین 0 و 1 است .مقدار 0 نشان دهنده یک مجموعه داده کاملاً خالص است، در حالی که مقدار 1 نشان دهنده یک مجموعه داده کاملاً ناپاک است.

Entropyمیزان اطلاعات یا عدم قطعیت در یک مجموعه داده را اندازه گیری می کند .مقدار entropy بین 0 و log(n) است، که n تعداد کلاس های موجود در مجموعه داده است .مقدار 0 نشان دهنده یک مجموعه داده کاملاً قطعی است، در حالی که مقدار log(n) نشان دهنده یک مجموعه داده کاملاً غیرقطعی است.

تفاوت اصلی بین Gini index و entropy در نحوه محاسبه آنها است Gini index .بر اساس توزیع احتمالات کلاس ها در یک مجموعه داده محاسبه می شود Entropy .بر اساس میزان اطلاعات یا عدم قطعیت در یک مجموعه داده محاسبه می شود.

- **Gini index** معمولاً برای مجموعه داده های با تعداد کلاس های زیاد استفاده می شود.
- **Entropy** معمولاً برای مجموعه داده های با تعداد کلاس های کم استفاده می شود.
- **Gini index** معمولاً برای مجموعه داده های با توزیع کلاس های نامتقارن استفاده می شود.
- **Entropy** معمولاً برای مجموعه داده های با توزیع کلاس های متقارن استفاده می شود.


Information gain=
1 − 0/94 = 0.09
entropy= 50%
-1,58
-0,58
Part
Full
none
some
entropy=0
entropy=0
yes
no
French
burger
thai
Italian
Fri
no
yes

مرحله دوم

predict(tree, example): این تابع برای پیش‌بینی مقدار خروجی بر اساس درخت تصمیم استفاده می‌شود. ورودی‌ها عبارتند از درخت تصمیم (tree) و مثال ورودی (example) که یک دیکشنری از ویژگی‌ها و مقادیر مربوط به مثال است. تابع با استفاده از درخت تصمیم به صورت بازگشتی مقدار پیش‌بینی شده را برمی‌گرداند.

makeInputValues(data): این تابع ویژگی‌های مختلف موجود در داده را جمع‌آوری کرده و مقادیر ممکن برای هر ویژگی را در attr_values ذخیره می‌کند

1. entropy(data): این تابع اندازه انتروپی (Entropy) را برای داده ورودی محاسبه می‌کند. انتروپی معیاری از انتشار نامی در داده است.
2. gini_index(data): این تابع معیار Gini Index را برای داده ورودی محاسبه می‌کند . Gini Index نشان‌دهنده خلوص داده است.
3. information_gain_entropy(data, attribute): این تابع میزان افزایش اطلاعات برای یک ویژگی مشخص را با استفاده از انتروپی محاسبه می‌کند.
4. gini_impurity(data, attribute): این تابع میزان تغییر در Gini Index را برای یک ویژگی مشخص محاسبه می‌کند.
5. information_gain_gini(data, attribute): این تابع میزان افزایش اطلاعات برای یک ویژگی مشخص را با استفاده از Gini Index محاسبه می‌کند.
6. select_best_attribute(data, inputs, criterion): این تابع به عنوان ورودی داده‌ها، ویژگی‌ها و معیار انتخاب ویژگی) معیار انتروپی یا (Gini Index را می‌گیرد و بهترین ویژگی بر اساس معیار انتخاب ویژگی انتخاب می‌کند.
7. CalcAccuracy(giniDecisionTree, entropyDecisionTree): این تابع دو درخت تصمیم برای معیارهای انتروپی و Gini Index را می‌گیرد و دقت پیش‌بینی‌های این درخت‌ها را محاسبه می‌کند.
8. BuildTree(data, inputs, criterion): این تابع یک درخت تصمیم را برای داده‌ها و ویژگی‌ها با استفاده از یک معیار مشخص) انتروپی یا (Gini Index ایجاد می‌کند.

9. Discreting(length, value, min, max): این تابع برای تبدیل اعداد پیوسته به اعداد گسسته با استفاده از تبدیل گسسته‌سازی فرکانسی (Discretization) استفاده می‌شود.

10. ReadFromCSV(filename): این تابع داده‌ها را از یک فایل CSV می‌خواند، آنها را پردازش می‌کند و آماده می‌کند تا برای ساخت درخت تصمیم استفاده شود.

```
Entropy accuracy :  83.8
Gini accuracy :  83.95
```

نتایج دقت برای داده تست بر روی درخت های ساخته شده توسط آنتروپی و جینی ایندکس

تعداد داده های train = 22000 (satisfied 11000 و neutral 11000)

الگوریتم های اصلی پروژه

1. ساخت درخت تصمیم :(BuildTree) این الگوریتم به منظور ساخت یک درخت تصمیم برای دسته‌بندی داده‌ها استفاده می‌شود. الگوریتم به صورت بازگشتی و با توجه به معیار انتخاب ویژگی) انتروپی یا (Gini Index و ویژگی‌های مختلف، برای هر گره درخت یک ویژگی را انتخاب می‌کند تا بهترین تقسیم برای داده‌ها را انجام دهد. این الگوریتم از توجه به انتروپی یا Gini Index برای محاسبه بهترین ویژگی استفاده می‌کند.

2.معیار انتخاب ویژگی (Entropy و Gini Index): درخت تصمیم برای تقسیم داده‌ها به ویژگی‌های مختلف نیاز دارد. معیار انتخاب ویژگی بر اساس این که چه ویژگی‌ای برای تقسیم بهتر است، از اهمیت بالایی برخوردار است. در این پروژه، دو معیار انتخاب ویژگی مورد استفاده قرار گرفته‌اند: انتروپی (Entropy) و معیار Gini Index. این معیارها برای اندازه‌گیری خلوص داده و تغییرات در داده‌ها برای ویژگی‌ها استفاده می‌شوند.

3.پیش‌بینی (Predict): بعد از ساخت درخت تصمیم، الگوریتم برای پیش‌بینی مقدار خروجی برای نمونه‌های جدید از درخت تصمیم استفاده می‌کند. با ورودی دادن نمونه به درخت، این الگوریتم با دنبال کردن مسیری در درخت تا رسیدن به یک برگ، مقدار پیش‌بینی شده را برمی‌گرداند.