DATA MINING FINAL REPORT

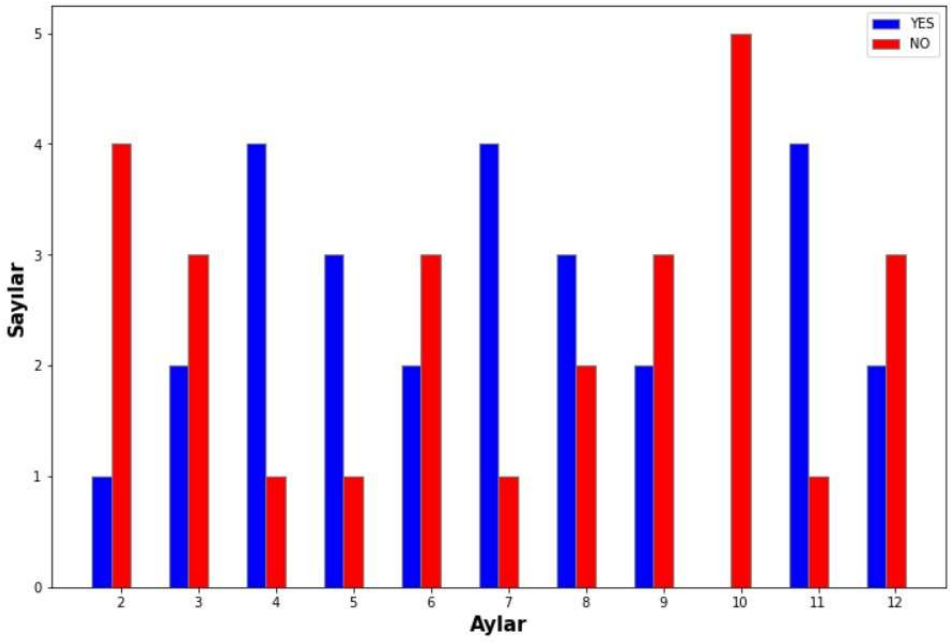
- Ravzanur CANTÜRK - 19120205029
- Selcen Fethiye MERSİNLİ - 19120205015
- Safiye Sena MERDİN - 18120205004

SORU: 2020 datasının aylara göre gruplanmış verilerinin "retail_and_recreation_percent_change_from_baseline", "parks_percent_change_from_baseline", "residential_percent_change_from_baseline" sütunlarının her aylarını negatif ve pozitif olarak classifice edin. Pozitif olma durumu insan yoğunluğunun belirtilen alanlarda fazla negatif olma durumu ise az olduğunu gösterir. Buna göre en az 5 kişilik anket yapınız ve çoğunluğun cevabına göre "go_out" classını oluşturunuz. Oluşan datayı train ve test data olarak ikiye ayırınız. Elde edilen bu dataları ve desicion tree methodunu kullanarak karar ağacı oluşturup dışarı çıkılıp çıkılmayacağına karar veriniz. En son accuracy hesaplatınız.

AMAÇ: 5 kişilik anketten karar ağacı oluşturarak sonraki kişilerin belirli koşullar altında dışarı çıkmak isteyip istemeyeceğinin tahmin edilmesi.

AYLARA GÖRE DIŞARI ÇIKMA/ÇIKMAMA TERCİH ANKETİ SONUÇLARI

Şubat	Ayı :	No	No	No	Yes	No
Mart	Ayı :	No	No	Yes	No	Yes
Nisan	Ayı :	No	Yes	Yes	Yes	Yes
Mayıs	Ayı :	Yes	No	Yes	No	Yes
Haziran	Ayı :	Yes	No	Yes	No	No
Temmuz	Ayı :	No	Yes	Yes	Yes	Yes
Ağustos	Ayı :	No	Yes	Yes	Yes	No
Eylül	Ayı :	Yes	No	No	Yes	No
Ekim	Ayı :	No	No	No	No	No
Kasım	Ayı :	Yes	Yes	Yes	No	Yes
Aralık	Ayı :	No	No	Yes	Yes	No



ÇÖZÜM:

- Öncelikle datadaki probleme uygun olmayan veriler çıkartılır.
- Data aylık olarak gruplanır.

date	retail_and_recreation_percent_change_from_baseline	parks_percent_change_from_baseline	residential_percent_change_from_baseline
2020-02-29	2.028130	4.509374	-0.698443
2020-03-31	-21.700010	-6.153226	6.332819
2020-04-30	-63.148836	-42.117354	20.742720
2020-05-31	-55.673896	-28.837900	17.455508
2020-06-30	-18.156210	32.502002	4.339668
2020-07-31	-6.876458	55.750799	-0.082189
2020-08-31	-9.014545	60.977582	-0.669343
2020-09-30	-13.071653	38.977984	2.389219
2020-10-31	-13.590351	30.077698	3.886670
2020-11-30	-26.801938	-1.853678	8.382888
2020-12-31	-47.541929	-26.385514	15.067144

- Datalar “positive” ve “negative” olarak kategorize edilir.

```
data
```

	retail_and_recreation_percent_change_from_baseline	parks_percent_change_from_baseline	residential_percent_change_from_baseline	go_out
0	RR-Positive	PP-Positive	R-Negative	NO
1	RR-Negative	PP-Negative	R-Positive	NO
2	RR-Negative	PP-Negative	R-Positive	YES
3	RR-Negative	PP-Negative	R-Positive	YES
4	RR-Negative	PP-Positive	R-Positive	NO
5	RR-Negative	PP-Positive	R-Negative	YES
6	RR-Negative	PP-Positive	R-Negative	YES
7	RR-Negative	PP-Positive	R-Positive	NO
8	RR-Negative	PP-Positive	R-Positive	NO
9	RR-Negative	PP-Negative	R-Positive	YES
10	RR-Negative	PP-Negative	R-Positive	NO

- Gruplanan datadan samplelar alınarak data split edilir (train(%54.54) ve test data(%45.46)).

```
train_data_m
```

	retail_and_recreation_percent_change_from_baseline	parks_percent_change_from_baseline	residential_percent_change_from_baseline	go_out
0	RR-Positive	PP-Positive	R-Negative	NO
5	RR-Negative	PP-Positive	R-Negative	YES
7	RR-Negative	PP-Positive	R-Positive	NO
9	RR-Negative	PP-Negative	R-Positive	YES
3	RR-Negative	PP-Negative	R-Positive	YES
8	RR-Negative	PP-Positive	R-Positive	NO

```
test_data_m
```

	retail_and_recreation_percent_change_from_baseline	parks_percent_change_from_baseline	residential_percent_change_from_baseline	go_out
1	RR-Negative	PP-Negative	R-Positive	NO
2	RR-Negative	PP-Negative	R-Positive	YES
4	RR-Negative	PP-Positive	R-Positive	NO
6	RR-Negative	PP-Positive	R-Negative	YES
10	RR-Negative	PP-Negative	R-Positive	NO

- Train datadan ID3 algoritması kullanılarak tree elde edilir.
- ID3 algoritması entropy ve information gain bilgilerini kullanarak decision tree oluşturmaya dayanır.

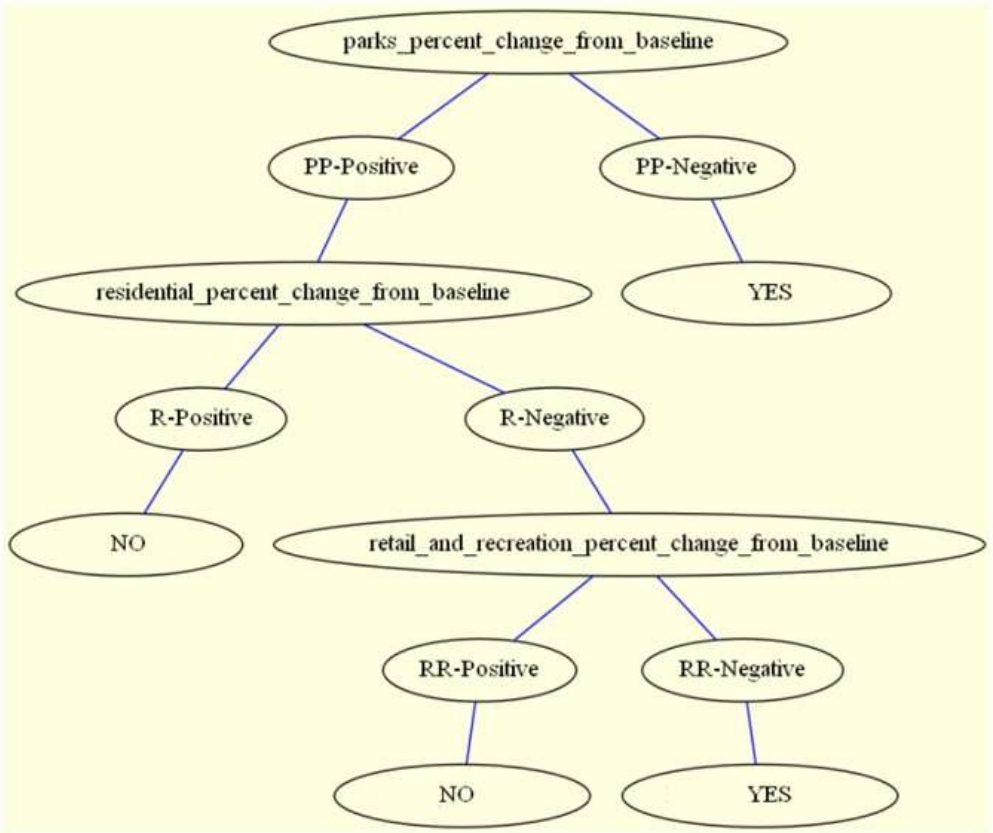
$$Entropy = -\sum_{i=0}^{c-1} p_i(t) log_2 p_i(t)$$

$$Gain_{split} = Entropy(p) - \sum_{i=1}^{k} \frac{n_i}{n} Entropy(i)$$

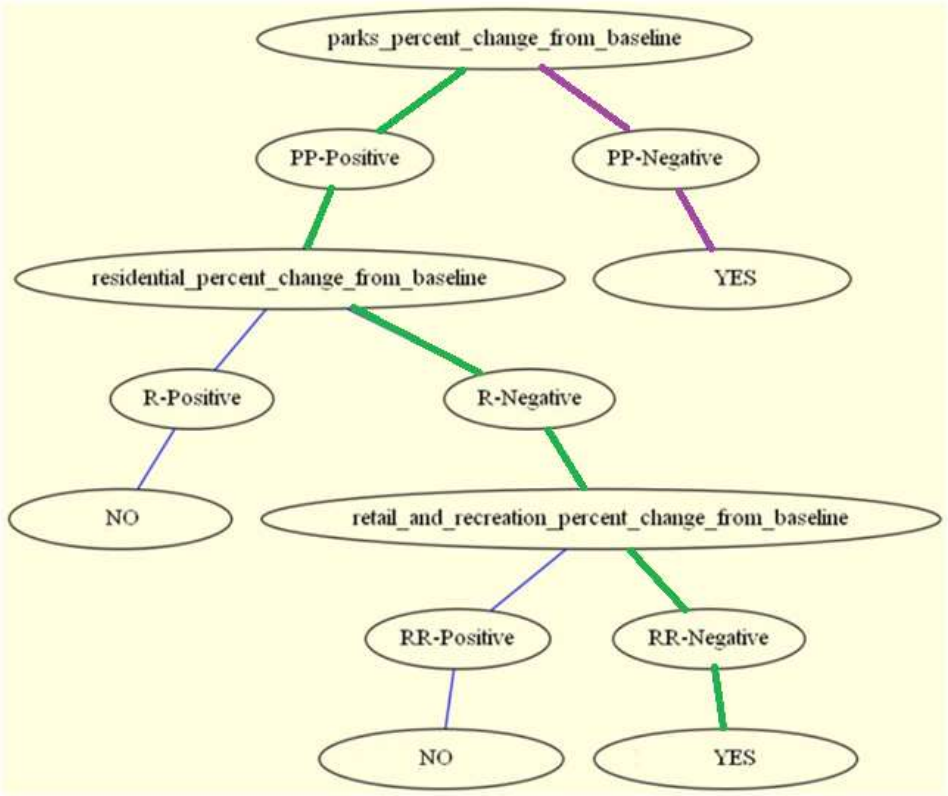
- Information gain değeri büyük olan feature, tree'ye node olarak eklenir ve eklenen node'un child değerleri belirlenir.
- Pure class seçilir. Feature value'ya karşılık gelen satırlar data set'ten çıkarılır.
- Data set boşalana kadar ve tüm sınıflar pure olana kadar aynı işlemler tekrarlanarak ağaç oluşturulur.

```
tree
```

```
{'parks_percent_change_from_baseline': {'PP-Positive': {'residential_percent_change_from_baseline': {'R-Positive': 'NO',
    'R-Negative': {'retail_and_recreation_percent_change_from_baseline': {'RR-Positive': 'NO',
      'RR-Negative': 'YES'}}}},
  'PP-Negative': 'YES'}}
```



- Oluşan ağaç üzerinden test datası kullanılarak tahmin yapılır.



Rule Set

R1: (Park=negative) ➔ YES

R2: (Park=positive, Residential=positive) ➔ NO

R3: (Park=positive, Residential=negative, Retail=positive) ➔ NO

R4: (Park=positive, Residential=negative, Retail=negative) ➔ YES

- Yapılan tahminler sonucunda accuracy hesaplanır.

```
accuracy
```

```
0.6
```

SONUÇ:

Oluşan tree, sonraki kişinin kararını %60 oranında doğru tahmin eder.