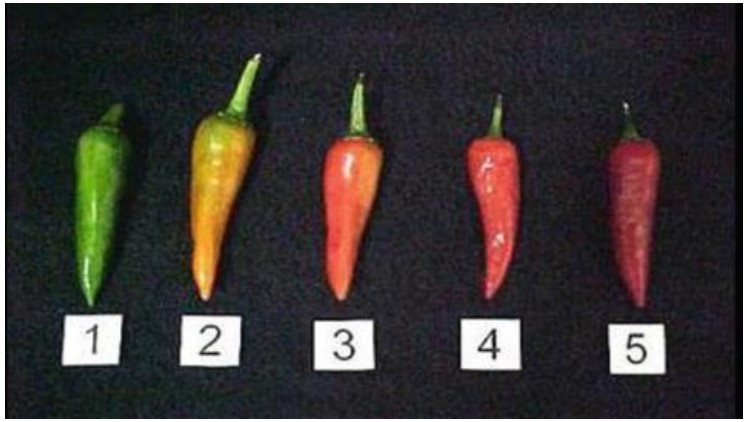
Možemo li koristiti kategorička obeležja?

Kategorička obeležja

Nominalna

Ordinalna



One-hot encoding

Ne postoji prirodan poredak

Ako dodelimo 1,2,3,… (*label encoding*) to utiče na računanje gubitka: model će podrazumevati poredak

One-hot-encoding

Br. komponenti = br. jedinstvenih obeležja

Kolona postaje vektor gde samo jedno obeležje ima vrednost 1

Label Encoding

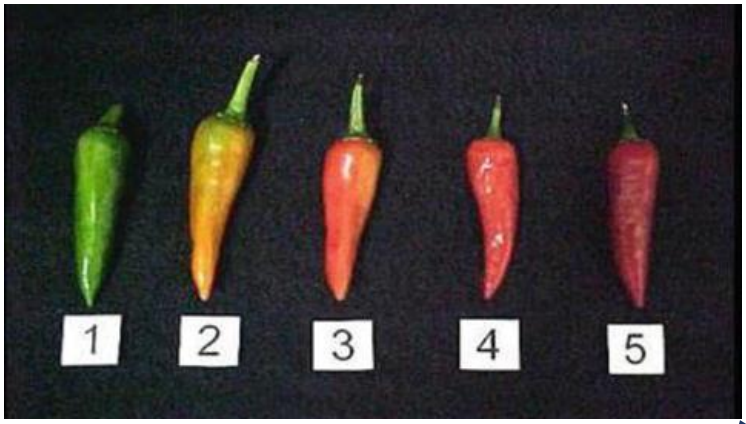
Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Možemo li koristiti kategorička obeležja?



Postoji prirodan poredak

Možemo li enkodirati sa 1,2,3,… ?

Zašto ne 100, 101 i 300 000?
Koliko je 2 jače od 1, a koliko je 3 jače od 2?

One-hot-encoding

https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02

Oprez – multikolinearnost

- Dva (ili više) obeležja su u korelaciji:

$$c_1 f_1 + c_2 f_2 + \cdots + c_k f_k = f_j$$

- Zašto je ovo problem?

$$y = \theta_0 + \theta_1 f_1 + \theta_2 f_2$$

$$f_1 = c_1 + c_2 f_2$$

Ali perfektna multikolinearnost se retko dešava. Možda onda nemamo problem?

- Ne možemo pronaći rešenje
  - Treba da procenimo uticaj f_1 na y dok držimo f_2 konstantnim
  - Ali f_2 se promeni kadgod se f_1 promeni!

Oprez – multikolinearnost

- Ipak je problematično…

- Teramo model da uči težinu koja nam ne treba (računarski resursi i vreme)
- Prokletstvo dimenzionalnosti
- Model je teže interpretirati
- θ su osetljive na specifičnosti trening skupa. Promeniće se drastično uklanjanjem varijabli

Multikolinearnost i *one-hot encoding*

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

$$Apple + Chicken + Broccoli = 1$$

Perfektna multikolinearnost:
Dummy Variable Trap

Rešenje: izbaciti jednu od tri varijable