4 - Bank - oplossingen

June 14, 2021

1. Lees het bestand bank.csv in. Het kan worden gebruikt om na te gaan of iemand een lening kan krijgen (kolom "pep" geeft dit weer)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 600 entries, 0 to 599
Data columns (total 12 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   id           600 non-null    object
 1   age          600 non-null    int64
 2   sex          600 non-null    object
 3   region       600 non-null    object
 4   income       600 non-null    float64
 5   married      600 non-null    object
 6   children     600 non-null    int64
 7   car          600 non-null    object
 8   save_act     600 non-null    object
 9   current_act  600 non-null    object
 10  mortgage     600 non-null    object
 11  pep          600 non-null    object
dtypes: float64(1), int64(2), object(9)
memory usage: 56.4+ KB
```

	id	age	sex	region	income	married	children	car	save_act	\
0	ID12101	48	FEMALE	INNER_CITY	17546.0	NO	1	NO	NO	
1	ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO	
2	ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES	
3	ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO	
4	ID12105	57	FEMALE	RURAL	50576.3	YES	0	NO	YES	

	current_act	mortgage	pep
0	NO	NO	YES
1	YES	YES	NO
2	YES	NO	NO
3	YES	NO	NO
4	NO	NO	NO

2. Dit bestand is niet direct bruikbaar voor ID3 en moet dus aangepast worden:

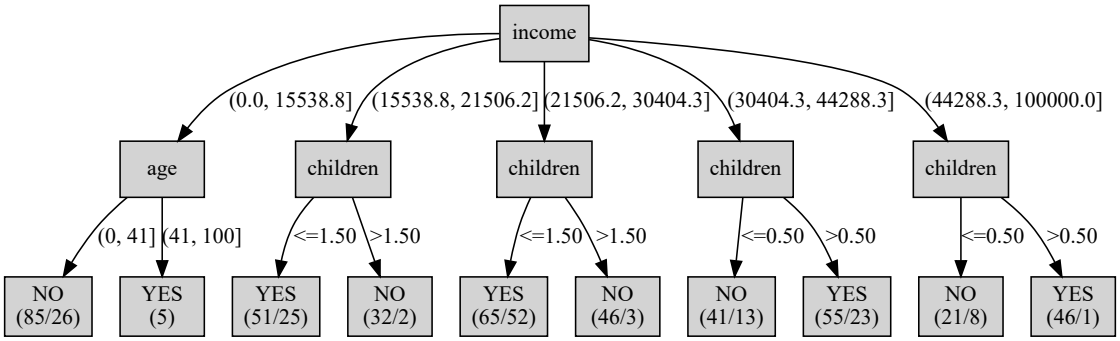
- de kolom "id" bevat unieke id's voor klanten en kunnen dus niets voorspellen. Verwijder deze

kolom.

- in dit bestand zitten ook continue variabelen (income, age, children)
- zet de waarden van “income” om in 5 categorieën (klassen)
-]0 - 15538,8]
-]15538,8 - 21506,2]
-]21506,2 - 30404,3]
-]30404,3 - 44288,3]
-]44288,3 - 100000]
- zet de waarden van “age” om in 2 categorieën:]0 - 41] en]41-100]
- de waarden van children kunnen maar 4 waarden aannemen, dus dat laten we zo

3. Maak een beslissingsboom met het ID3 algoritme. Probeer de resulterende boom te interpreteren. Komt dit overeen met je intuïtie?

```
Id3Estimator(gain_ratio=False, is_repeating=False, max_depth=2,
             min_entropy_decrease=0.0, min_samples_split=2, prune=False)
```

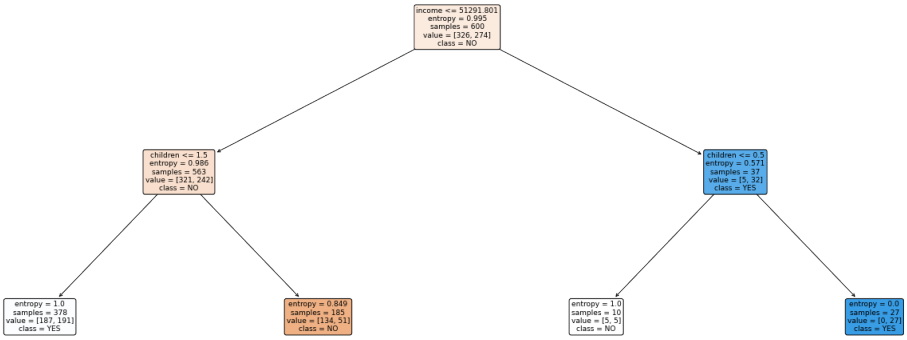



```
0.745
```

4. Lees de oorspronkelijke data opnieuw in. Verwijder alle nominale en ordinale attributen. Aangezien er continue variabelen zijn, kan je niet kiezen voor ID3 algoritme, maar je kan wel kiezen voor C4.5 m.b.v. de Scikit Learn DecisionTreeClassifier. Voer dit algoritme uit en kijk wat het resultaat is. Is dit beter of slechter dan ID3?

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='entropy',
                       max_depth=2, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=None, splitter='best')
```

```
0.595
```



5. Lees de oorspronkelijke data opnieuw in. Ditmaal gaan geen attributen verwijderen en toch C4.5 gebruiken m.b.v. de Scikit Learn DecisionTreeClassifier en Pandas `get_dummies`. Voer dit algoritme uit en kijk wat het resultaat is. Is dit beter of slechter dan vorige oplossingen?

```
DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='entropy',
                       max_depth=2, max_features=None, max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort='deprecated',
                       random_state=None, splitter='best')
```

```
0.595
```