Veri Madenciliğine Giriş Dönem Projesi

Big Five Kişilik Testi Analizi

Cengizhan PARLAK

Big Five Personality Traits

- OCEAN kısaltmasıyla da gösterilebilen, Türkçe'ye Beş Büyük Faktör Kuramı olarak çevrilen Beş Büyük kişilik özelliği (Big Five Personality Traits), kişilik özellikleri için bir taksonomi veya gruplamadır. OCEAN kısaltmasındaki harfler:
- openness: yeni deneyim ve fikirlere açıklık
- conscientiousness: yapılan işe, kedine ve çevreye karşı sorumluluk hissi; düzenli ve hazırlıklı olma
- extroversion: insanlarla olan ilişkilerde dışadönüklük
- agreeablenes: farklı durumlara ve ortama uyum sağlayabilme, onaylanma
- neuroticism: öfke, endişe, korku, imrenme vb. duyguları yoğun olarak yaşama, nevrotiklik

Veri Seti

- Veri setinde: bahsedilen 5 karakter özelliğiyle ilgili toplamda 50 soruya verilen cevaplar yer almaktadır.
- Teste katılan kişi cevaplarında, soruda bahsedilen duruma uyup uymadığını belirtir.
- «1» cevabı: «katılmıyorum/beni anlatmıyor» anlamındayken;
- «5» cevabı: «katılıyorum/beni anlatıyor» anlamındadır.
- Herhangi bir karakter özelliğindeki 10 soruya verilen cevabın ortalamasının yüksekliği; o özelliğin baskın olarak kişide olduğunu belirtir.
- Dışadönüklük sorularına verdiği cevapların ortalaması 4,2 olan birisinin dışadönük bir karakterde olması beklenir.
- Veri setinde: teste katılan kişilerin ülkeleri; soruları ve testi bitirme süreleri gibi bilgiler de yer almaktadır.
- Veri seti 1 milyonun üzerinde giriş içermektedir.

```
Katılımcı sayısı:  1015341
```

Out[4]:

	EXT1	EXT2	EXT3	EXT4	EXT5	EXT6	EXT7	EXT8	EXT9	EXT10	...	dateload	screenw	screenh	introelapse	testelapse	endelapse	IPC	country
0	4.0	1.0	5.0	2.0	5.0	1.0	5.0	2.0	4.0	1.0	...	2016-03-03 02:01:01	768.0	1024.0	9.0	234.0	6	1	GB
1	3.0	5.0	3.0	4.0	3.0	3.0	2.0	5.0	1.0	5.0	...	2016-03-03 02:01:20	1360.0	768.0	12.0	179.0	11	1	MY
2	2.0	3.0	4.0	4.0	3.0	2.0	1.0	3.0	2.0	5.0	...	2016-03-03 02:01:56	1366.0	768.0	3.0	186.0	7	1	GB
3	2.0	2.0	2.0	3.0	4.0	2.0	2.0	4.0	1.0	4.0	...	2016-03-03 02:02:02	1920.0	1200.0	186.0	219.0	7	1	GB
4	3.0	3.0	3.0	3.0	5.0	3.0	3.0	5.0	3.0	4.0	...	2016-03-03	1366.0	768.0	8.0	315.0	17	2	KE

Veri seti yapısı ve boyutu

Dışadönüklük ve Nevrotiklik ile İlgili Sorular

Code	Question	Key		Code	Question	Key
EXT1	I am the life of the party.	(+)	|	EST1	I get stressed out easily.	(+)
EXT2	I don't talk a lot.	(-)	|	EST2	I am relaxed most of the time.	(-)
EXT3	I feel comfortable around people.	(+)	|	EST3	I worry about things.	(+)
EXT4	I keep in the background.	(-)	|	EST4	I seldom feel blue.	(-)
EXT5	I start conversations.	(+)	|	EST5	I am easily disturbed.	(+)
EXT6	I have little to say.	(-)	|	EST6	I get upset easily.	(+)
EXT7	I talk to a lot of different people at parties.	(+)	|	EST7	I change my mood a lot.	(+)
EXT8	I don't like to draw attention to myself.	(-)	|	EST8	I have frequent mood swings.	(+)
EXT9	I don't mind being the center of attention.	(+)	|	EST9	I get irritated easily.	(+)
EXT10	I am quiet around strangers.	(-)	|	EST10	I often feel blue.	(+)

Uyum ve Sorumluluk ile İlgili Sorular

Code	Question	Key		Code	Question	Key
AGR1	I feel little concern for others.	(-)	|	CSN1	I am always prepared.	(+)
AGR2	I am interested in people.	(+)	|	CSN2	I leave my belongings around.	(-)
AGR3	I insult people.	(-)	|	CSN3	I pay attention to details.	(+)
AGR4	I sympathize with others' feelings.	(+)	|	CSN4	I make a mess of things.	(-)
AGR5	I am not interested in other people's problems.	(-)	|	CSN5	I get chores done right away.	(+)
AGR6	I have a soft heart.	(+)	|	CSN6	I often forget to put things back in their proper place.	(-)
AGR7	I am not really interested in others.	(-)	|	CSN7	I like order.	(+)
AGR8	I take time out for others.	(+)	|	CSN8	I shirk my duties.	(-)
AGR9	I feel others' emotions.	(+)	|	CSN9	I follow a schedule.	(+)
AGR10	I make people feel at ease.	(+)	|	CSN10	I am exacting in my work.	(+)

Deneyime Açıklık ile İlgili Sorular

Code	Question	Key
OPN1	I have a rich vocabulary.	(+)
OPN2	I have difficulty understanding abstract ideas.	(-)
OPN3	I have a vivid imagination.	(+)
OPN4	I am not interested in abstract ideas.	(-)
OPN5	I have excellent ideas.	(+)
OPN6	I do not have a good imagination.	(-)
OPN7	I am quick to understand things.	(+)
OPN8	I use difficult words.	(-)
OPN9	I spend time reflecting on things.	(+)
OPN10	I am full of ideas.	(+)

Problem Tanımı

- Veri setinde:
  - 50 soruya verilen cevaplar, soruların cevaplanma süreleri; katılımcıların ülkeleri ve testi yaptıkları cihazın ekran çözünürlükleri gibi bilgilerle birlikte testin yapılma zamanı gibi veriler yer almaktadır.
  - herhangi bir hedef değişken bulunmamaktadır.
  - problem tanımı olarak: 50 soruya verilen cevap(lar) üzerinden, katılımcının ülkesinin hangi doğrulukta tahmin edilebildiğinin çalışması yapılmıştır.

Uygulanan Veri Madenciliği Algoritmaları

- Veri setindeki sınıflandırmayı gerçekleştirmek için makine öğrenme algoritmalarından Lojistik Regresyon (Logistic Regression) ve Karar Ağacı (Decision Tree) kullanılmıştır.
- Çok çıktılı problemlerin çözümünde kullanılabilen Karar Ağacı algoritması; verilen cevapların sayısal değeri üzerinden sınıflandırma yapmak için problem tanımına uygun gözükmüştür.
- Lojistik Regresyon ise ikili (non-binary) olmayan tahminlerde «biri ve de diğerleri» (one-vs-all) metodu kullanılarak uygulanabilirliği sayesinde tercih edilmiştir.

Veri İşlemleri

- 50 soru içerisinde cevap olarak olmaması gereken «0» değerlerinin silinmesi:

```
In [7]: #cevabı 0 olan soru olmaması gerektiği halde veri setinde 0 değerleri var. bu hatalı girişlerin silinmesi gerekiyor
        data_set = data_set[(data_set[answer_columns] != 0).all(axis = 1)]
```

- Herhangi «null» değer içeren satırların veri setinden silinmesi:

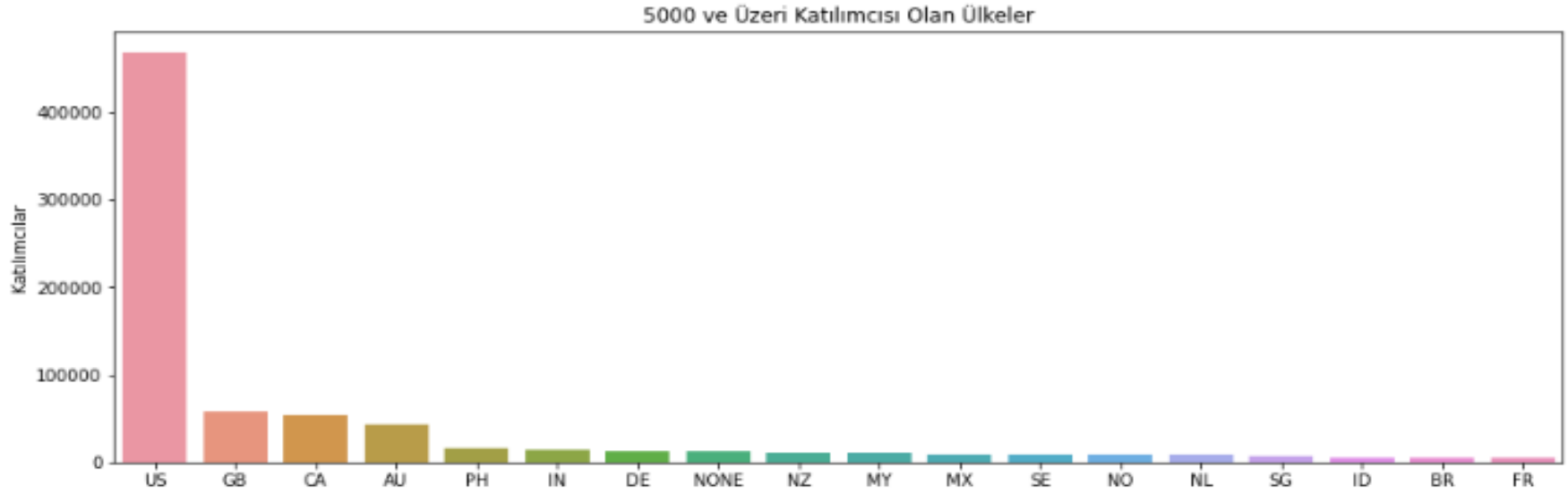
```
In [5]: print('Eksik değeler var mı? ', data_set.isnull().values.any())
        print('Ne kadar eksik değer var? ', data_set.isnull().values.sum())
        data_set.dropna(inplace=True)
        print('Eksik değerlerin olduğu satırları sildikten sonraki katılımcı sayısı: ', len(data_set))

        Eksik değeler var mı?  True
        Ne kadar eksik değer var?  186358
        Eksik değerlerin olduğu satırları sildikten sonraki katılımcı sayısı:  1012050
```

Veri İşlemleri

- Katılımcıların ülkeleri:

```
In [10]: # Katılımcıların ülke dağılımları
countries = pd.DataFrame(data_set['country'].value_counts())
countries_5000 = countries[countries['country'] >= 5000]
plt.figure(figsize=(15,5))
sns.barplot(data=countries_5000, x=countries_5000.index, y='country')
plt.title('5000 ve Üzeri Katılımcısı Olan Ülkeler')
plt.ylabel('Katılımcılar');
```



Algoritma Adımları

- Cevap sütunları olarak ilk 50 sütunun, veri girişi olarak tüm satırların alınması;

```
In [30]: answer_data = data_set.iloc[:,0:50]
```

- Veri setindeki ülke değerlerinin alınması:

```
In [31]: answer_data['country'] = data_set['country']
```

- Korelasyon değerlerinin seçilmesi:

```
In [32]: for col in answer_data.columns:
             answer_data[col] = answer_data[col].astype('category').cat.codes

In [33]: corr_data = pd.DataFrame(answer_data.corr()['country'][:])

In [34]: corr_data = corr_data.reset_index()

In [35]: top_correlation = corr_data.sort_values('country', ascending=False).head(10)['index'].to_list()

In [36]: least_correlation = corr_data.sort_values('country', ascending=False).tail(5)['index'].to_list()

In [37]: correlation_data = answer_data[top_correlation+least_correlation]

In [38]: target_data = answer_data['country']
```

Algoritma Adımları ve Sonuçlar

- Bağıntı verileri, hedef veri ve test büyüklüğünün seçilmesi ve modelin eğitilmesi:

```
In [39]: var_train, var_test, res_train, res_test = train_test_split(correlation_data, target_data, test_size = 0.3)

In [40]: logistic_reg = LogisticRegression(random_state=0).fit(var_train, res_train)
```

- Lojistik Regresyon ile veri setindeki doğruluk oranı:

```
In [41]: prediction = logistic_reg.predict(var_test)

In [42]: accuracy_score(res_test, prediction)
Out[42]: 0.7140735707305155
```

- Karar Ağacı algoritması ile ortaya çıkan doğruluk oranı:

```
In [43]: decision_tree = tree.DecisionTreeClassifier()
         decision_tree = decision_tree.fit(var_train, res_train)

In [44]: decision_prediction = decision_tree.predict(var_test)

In [45]: accuracy_score(res_test, decision_prediction)
Out[45]: 0.999965642560469
```

Algoritma Adımları ve Sonuçlar

US	GB	CA	AU
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

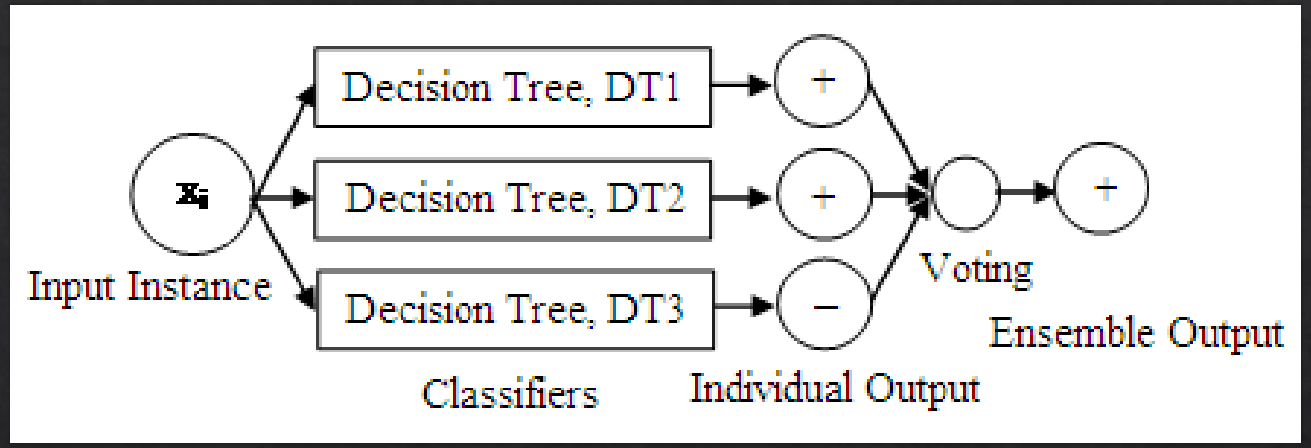
One-vs-all (one-vs-rest) ya da «biri ve diğerleri» metodu ile Lojistik Regresyon algoritmasının, iki değişkenli olmayan verilerde uygulanması şekilde verildiği gibi gerçekleşmektedir.

- Modelin eğitilmesinden sonra, her giriş için: tahmin edilen ülke değeri 1; geri kalan ülke değerleri 0 yapılır.
- Tahmin işleminde, katılımcıların ilk 50 soruya verdiği cevaplara bakılmasının ardından: katılımcının gerçek ülkesi ile, 1 olarak yazılmış: 'tahmin edilen' ülkesi karşılaştırılır.
- Model, 0.71 başarı oranıyla katılımcının ülkesini tahmin etmektedir.

Algoritma Adımları ve Sonuçlar

Karar Ağacı algoritmasında 50 soruya verilen cevaplara göre ağaç yapısı oluşturulur ve eğitilir.

Sorulara verilen integer (tamsayı) tipindeki cevapların kıyaslanması ile ülke tahmini yapılır.



- Tahmin edilecek ülke değeri için, sorulara verilen cevaplara göre ağaç üzerinde değerlendirme yapılır.
- Sorulara verilen cevapların, test verisindeki cevaplara benzerliği üzerinden dallanarak ülke tahmini gerçekleştirilir.
- Ağaç yapısı sonunda tahmin edilen ülke, katılımcının gerçek ülkesiyle karşılaştırılır.
- Model, 0.99 başarı oranıyla katılımcının ülkesini tahmin etmektedir.