



İSTANBUL MEDENİYET ÜNİVERSİTESİ
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

Grup Üyeleri:

Teslim tarihi: 09.01.2022

Fatima Ali 18120205042

Abdulbari Saka 18120212008

Data Mining Final Projesi Raporu

Özet

Veri setinde sub_region_1 sütununda verilen bulunan farklı mekânlar, veri ile verilen tarihteki peak yoğunluğunun pandemi öncesi değerlere(baseline) göre değişim yüzdesi gösterilmektedir (yani ziyaretçi sayısındaki değişim).

Veri setinde bulunan sütunlar, country_region_code , country_region , sub_region_1 , sub_region_2 , metro_area , iso_3166_2_code , census_fips_code , place_id , date , retail_and_recreation_percent_change_from_baseline , grocery_and_pharmacy_percent_change_from_baseline , parks_percent_change_from_baseline ,transit_stations_percent_change_from_baseline ,

workplaces_percent_change_from_baseline ,residential_percent_change_from_baseline

Toplanılmış bütün bilgiler dataset20 ve dataset21 veri setlerinde yerleştirilmiş. Ama bazı bilgiler onun karşısında bilinmeyen bilgiler vardır şehri olamayan bilgiler gibidir. Analiz doğru bir şekilde yapabilmek için eksik bilgiler bulunması veya tahmin edilmesi gerek. Bu işlemleri gerçekleştirmek için belirli metotlar uygulandı. Bunlar bu raporun içerisinde açıklandı.

Problem

2021 ve 2020 covid19 veri setinde içerisinde şehri (sub_region_1) belli olmayan veriler vardır NaN/undefined olarak adlandırılıyor. Bu tip veriler içirin ver seti her hangi bir işlem yapmaya kalkarsak doğru sonuç alamayabiliriz.

	sub_region_1	retail_and_recreation_percent_change_from_baseline	grocery_and_pharmacy_percent_change_from_baseline	parks_percent_change_from_baseline
0	undefined	-88.0	-64.0	-71
1	undefined	-86.0	-57.0	-71
2	undefined	-85.0	-57.0	-66
3	undefined	-33.0	18.0	4
4	undefined	-38.0	12.0	-8
...
273	undefined	10.0	63.0	34
274	undefined	8.0	50.0	26
275	undefined	9.0	42.0	53
276	undefined	16.0	61.0	39
277	undefined	13.0	62.0	33

278 rows x 5 columns

Amaç

Amacımız Nan veya undefined veriler hangi şehre (sub_region_1) ait olduğuna bulmak.

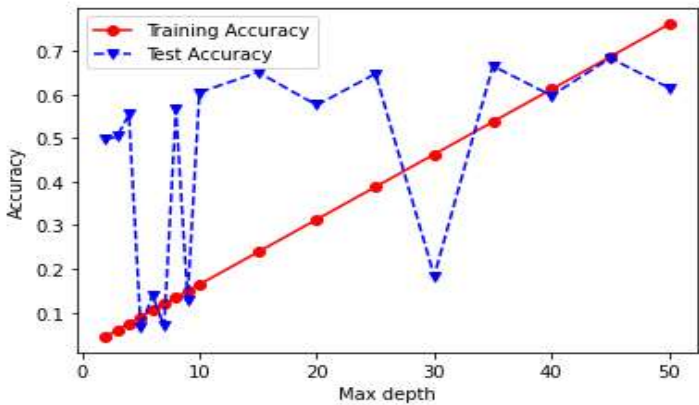
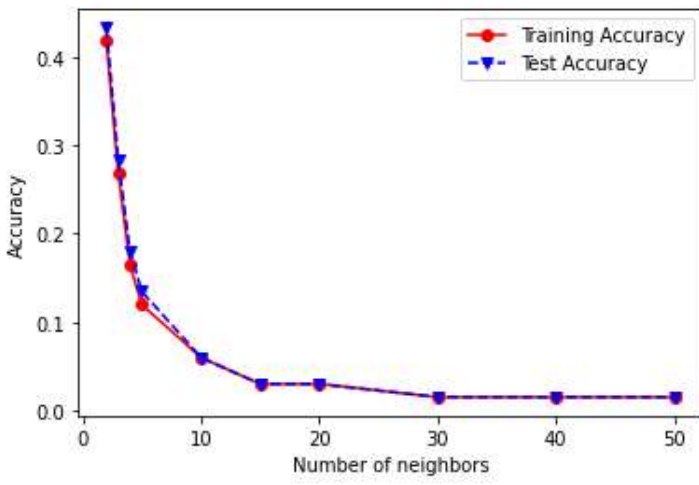
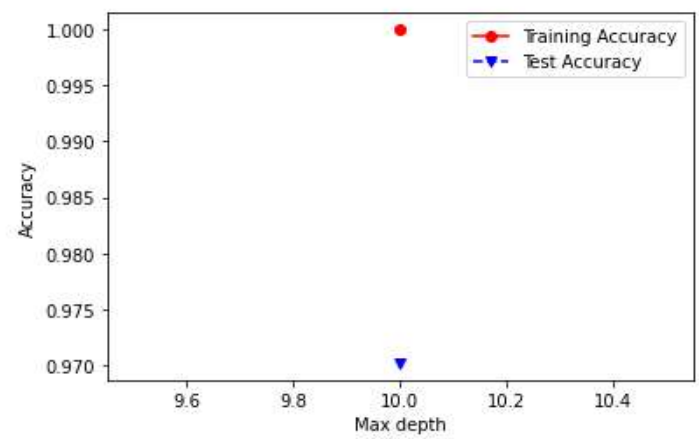
	sub_region_1	retail_and_recreation_percent_change_from_baseline	grocery_and_pharmacy_percent_change_from_baseline	parks_percent_change_from_baseline
0	Kars	-88.0	-64.0	-71
1	Kars	-86.0	-57.0	-71
2	Kars	-85.0	-57.0	-66
3	Kocaeli	-33.0	18.0	4
4	Kirkkale	-38.0	-12.0	-8
...
273	Yozgat	10.0	63.0	34
274	Gaziantep	8.0	50.0	26
275	Yozgat	9.0	42.0	53
276	Yozgat	16.0	61.0	38
277	Yozgat	13.0	62.0	33

Yöntemler

Elimizde bulunan veri setinde bilgiler çok olunca veya benzerlik olmayınca oluşturduğumuz modeller doğru bir şekilde geliştirdiğimiz tespit etmek için veri setiler farklı iki yöntem işlemler ayırdık.

- **İlk yöntem**

Modelimiz doğru bir şekilde geliştirdiğimiz öğrenebilmek için temel veri setinden sub_region_1 groupby grupladık ve ortalamasına aldıktan sonra testDataseti için sample aldık. Bu şekilde bir model elde ettik.

Methode	Train Accuracy	Test Accuracy	Plot
Decision Tree Classifier	0.69	0.69	
K-Nearest neighbor classifier	0.45	0.47	
Ensemble Methods Random Forest Classifier	1.00	0.97	

- İlk Yöntemin Yorumu

Üç tane metot kullanıldı her birinden farklı sonuçlar elde ettik

Ama en iyi metot seçmek gerekirse Random Forest Classifier bu tip veri set için en iyi tahminler sağlayacak.

- İkinci yöntem

Bu kısımda kullandığımız veri seti fazla karmaşıklığı yok çünkü veri setisi ikiye böldük test için kısım aldık kalan kısımlar train için kullanıldı. Üç tane metot kullanıldı her birinden farklı sonuçlar elde ettik.

Method	Train Accuracy	Test Accuracy	Plot																																																						
Decision Tree Classifier	0.98	0.92	<p>Plot showing Accuracy vs Max depth for Decision Tree Classifier. Training Accuracy (red solid line with circles) and Test Accuracy (blue dashed line with triangles) both increase with Max depth, reaching a plateau of 1.0 at depth 30.</p> <table><thead><tr><th>Max depth</th><th>Training Accuracy</th><th>Test Accuracy</th></tr></thead><tbody><tr><td>0</td><td>0.18</td><td>0.02</td></tr><tr><td>2</td><td>0.20</td><td>0.05</td></tr><tr><td>4</td><td>0.22</td><td>0.08</td></tr><tr><td>6</td><td>0.25</td><td>0.12</td></tr><tr><td>8</td><td>0.28</td><td>0.18</td></tr><tr><td>10</td><td>0.35</td><td>0.25</td></tr><tr><td>12</td><td>0.45</td><td>0.35</td></tr><tr><td>14</td><td>0.58</td><td>0.48</td></tr><tr><td>16</td><td>0.75</td><td>0.65</td></tr><tr><td>18</td><td>0.85</td><td>0.80</td></tr><tr><td>20</td><td>0.90</td><td>0.85</td></tr><tr><td>25</td><td>0.98</td><td>0.98</td></tr><tr><td>30</td><td>1.00</td><td>1.00</td></tr><tr><td>35</td><td>1.00</td><td>1.00</td></tr><tr><td>40</td><td>1.00</td><td>1.00</td></tr><tr><td>45</td><td>1.00</td><td>1.00</td></tr><tr><td>50</td><td>1.00</td><td>1.00</td></tr></tbody></table>	Max depth	Training Accuracy	Test Accuracy	0	0.18	0.02	2	0.20	0.05	4	0.22	0.08	6	0.25	0.12	8	0.28	0.18	10	0.35	0.25	12	0.45	0.35	14	0.58	0.48	16	0.75	0.65	18	0.85	0.80	20	0.90	0.85	25	0.98	0.98	30	1.00	1.00	35	1.00	1.00	40	1.00	1.00	45	1.00	1.00	50	1.00	1.00
Max depth	Training Accuracy	Test Accuracy																																																							
0	0.18	0.02																																																							
2	0.20	0.05																																																							
4	0.22	0.08																																																							
6	0.25	0.12																																																							
8	0.28	0.18																																																							
10	0.35	0.25																																																							
12	0.45	0.35																																																							
14	0.58	0.48																																																							
16	0.75	0.65																																																							
18	0.85	0.80																																																							
20	0.90	0.85																																																							
25	0.98	0.98																																																							
30	1.00	1.00																																																							
35	1.00	1.00																																																							
40	1.00	1.00																																																							
45	1.00	1.00																																																							
50	1.00	1.00																																																							
K-Nearest neighbor classifier	0.75	0.75	<p>Plot showing Accuracy vs Number of neighbors for K-Nearest neighbor classifier. Training Accuracy (red solid line with circles) and Test Accuracy (blue dashed line with triangles) both decrease as the number of neighbors increases, with Test Accuracy dropping more sharply after 4 neighbors.</p> <table><thead><tr><th>Number of neighbors</th><th>Training Accuracy</th><th>Test Accuracy</th></tr></thead><tbody><tr><td>1</td><td>1.00</td><td>1.00</td></tr><tr><td>2</td><td>0.75</td><td>0.75</td></tr><tr><td>3</td><td>0.70</td><td>0.68</td></tr><tr><td>4</td><td>0.68</td><td>0.65</td></tr><tr><td>5</td><td>0.66</td><td>0.60</td></tr><tr><td>6</td><td>0.65</td><td>0.58</td></tr><tr><td>7</td><td>0.64</td><td>0.56</td></tr><tr><td>8</td><td>0.63</td><td>0.53</td></tr><tr><td>9</td><td>0.62</td><td>0.50</td></tr><tr><td>10</td><td>0.61</td><td>0.49</td></tr></tbody></table>	Number of neighbors	Training Accuracy	Test Accuracy	1	1.00	1.00	2	0.75	0.75	3	0.70	0.68	4	0.68	0.65	5	0.66	0.60	6	0.65	0.58	7	0.64	0.56	8	0.63	0.53	9	0.62	0.50	10	0.61	0.49																					
Number of neighbors	Training Accuracy	Test Accuracy																																																							
1	1.00	1.00																																																							
2	0.75	0.75																																																							
3	0.70	0.68																																																							
4	0.68	0.65																																																							
5	0.66	0.60																																																							
6	0.65	0.58																																																							
7	0.64	0.56																																																							
8	0.63	0.53																																																							
9	0.62	0.50																																																							
10	0.61	0.49																																																							
Ensemble Methods Random Forest Classifier	1.00	1.00	<p>Plot showing Accuracy vs Max depth for Ensemble Methods Random Forest Classifier. Training Accuracy (red solid line with circles) and Test Accuracy (blue dashed line with triangles) are both 1.00 across the range of Max depth from 9.6 to 10.4.</p> <table><thead><tr><th>Max depth</th><th>Training Accuracy</th><th>Test Accuracy</th></tr></thead><tbody><tr><td>9.6</td><td>1.00</td><td>1.00</td></tr><tr><td>9.8</td><td>1.00</td><td>1.00</td></tr><tr><td>10.0</td><td>1.00</td><td>1.00</td></tr><tr><td>10.2</td><td>1.00</td><td>1.00</td></tr><tr><td>10.4</td><td>1.00</td><td>1.00</td></tr></tbody></table>	Max depth	Training Accuracy	Test Accuracy	9.6	1.00	1.00	9.8	1.00	1.00	10.0	1.00	1.00	10.2	1.00	1.00	10.4	1.00	1.00																																				
Max depth	Training Accuracy	Test Accuracy																																																							
9.6	1.00	1.00																																																							
9.8	1.00	1.00																																																							
10.0	1.00	1.00																																																							
10.2	1.00	1.00																																																							
10.4	1.00	1.00																																																							

- **İkinci Yöntemin Yorumu**

En iyi metod seçmek gerekirse Decision Tree Classifier bu tip veri seti için en iyi tahminler sağlayacak.

Sonuç

Geliştirdiğimiz modellerin doğru bir şekilde çalıştığından emin olmak için veri seti değiştirmek için iki tane yöntem kullandık. Her yöntemde üç tane sınıflandırıcı metodu kullandık. Her metottan farklı sonuçlar elde ettik. Bazıları yüzde yüz test ve train accuracy sonuçları verdi. Ama böyle tip veri seti 75% - 90 % arasında sonuçlar bekliyoruz. Aynı anda ikinci yöntem mantıklı olduğuna söyleyebiliriz çünkü her hangi bir model geliştirirken veriseti ikiye bölünmesi gerekiyor bir kısım train için diğer kısım test için olmalı.