



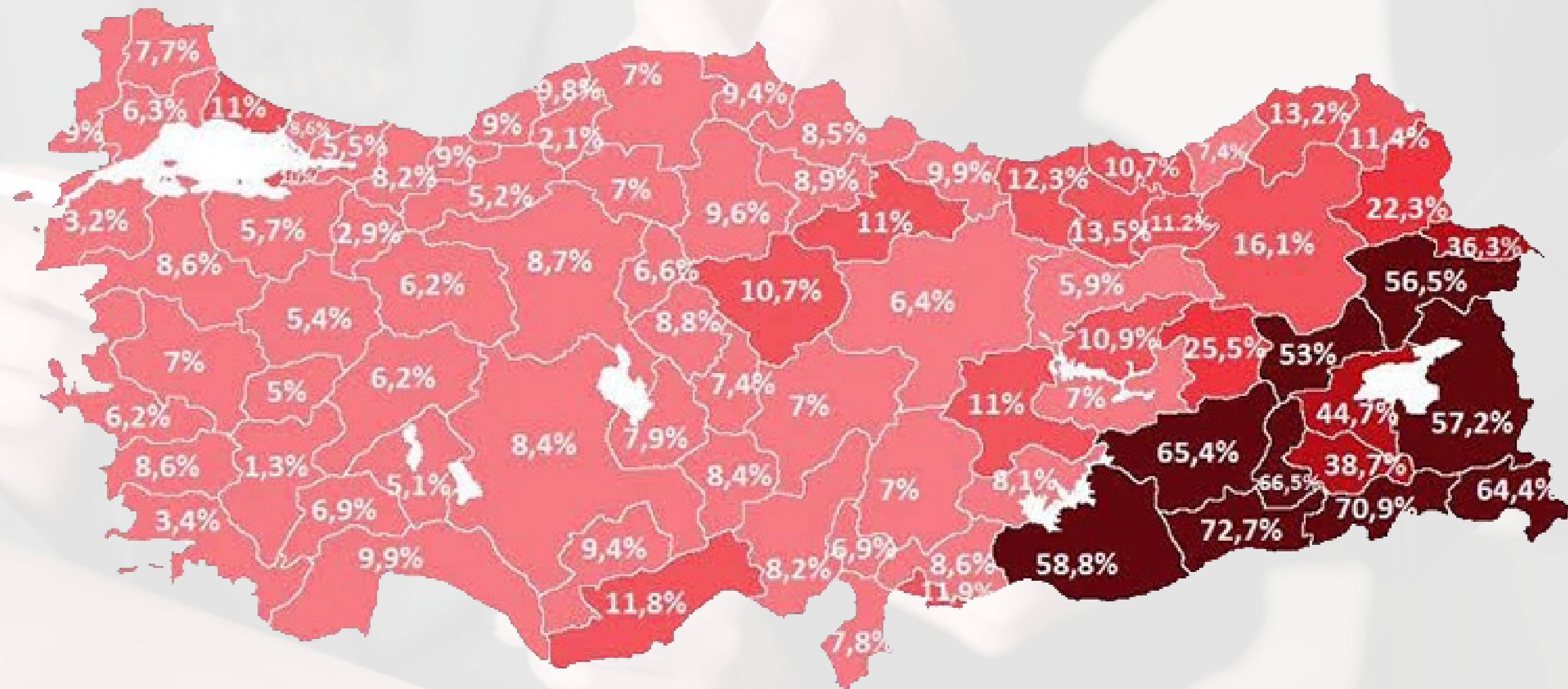
Patika.dev & EnerjiSA Veri Bilimi ve Analitiği Bootcamp

Nitelikli Kaçak
Tahminleme Projesi

KULLANIM HARİTASI

The map displays the usage percentage of a service across the provinces of Turkey. The data is as follows:

Province	Usage Percentage
İstanbul	11%
Trakya	7,7%
Edirne	9%
Yalova	6,3%
İzmir	11%
Manisa	8,6%
Denizli	5,5%
Uşak	8,2%
Bilecik	9%
Bursa	5,2%
Yozgat	9,8%
Çankırı	7%
Ordu	9,4%
Samsun	8,5%
Trabzon	8,9%
Rize	9,9%
Giresun	12,3%
Erzurum	10,7%
Van	7,4%
Siirt	13,2%
Bitlis	11,4%
Mardin	22,3%
Hakkâri	36,3%
Şırnak	56,5%
Diyarbakır	25,5%
Gaziantep	53%
Adana	44,7%
İzmir	57,2%
Denizli	64,4%
Manisa	70,9%
Uşak	66,5%
Bilecik	65,4%
Bursa	72,7%
Yozgat	58,8%
Çankırı	8,1%
Ordu	11%
Samsun	7%
Rize	10,9%
Giresun	7%
Erzurum	11%
Van	10,9%
Siirt	5,9%
Bitlis	6,4%
Mardin	11%
Hakkâri	10,7%
Şırnak	8,8%
Diyarbakır	6,6%
Gaziantep	8,7%
Adana	6,2%
İzmir	5,4%
Denizli	5%
Manisa	6,2%
Uşak	7%
Bilecik	6,2%
Bursa	8,6%
Yozgat	1,3%
Çankırı	8,6%
Ordu	3,4%
Samsun	6,9%
Rize	5,1%
Giresun	9,9%
Erzurum	8,4%
Van	7,9%
Siirt	7,4%
Bitlis	8,4%
Mardin	7%
Hakkâri	8,2%
Şırnak	6,9%
Diyarbakır	11,8%
Gaziantep	9,4%
Adana	8,2%
İzmir	11,9%
Denizli	7,8%



Yol Haritası

1



Değişken
Analizi

2



Veri
Temizleme

3

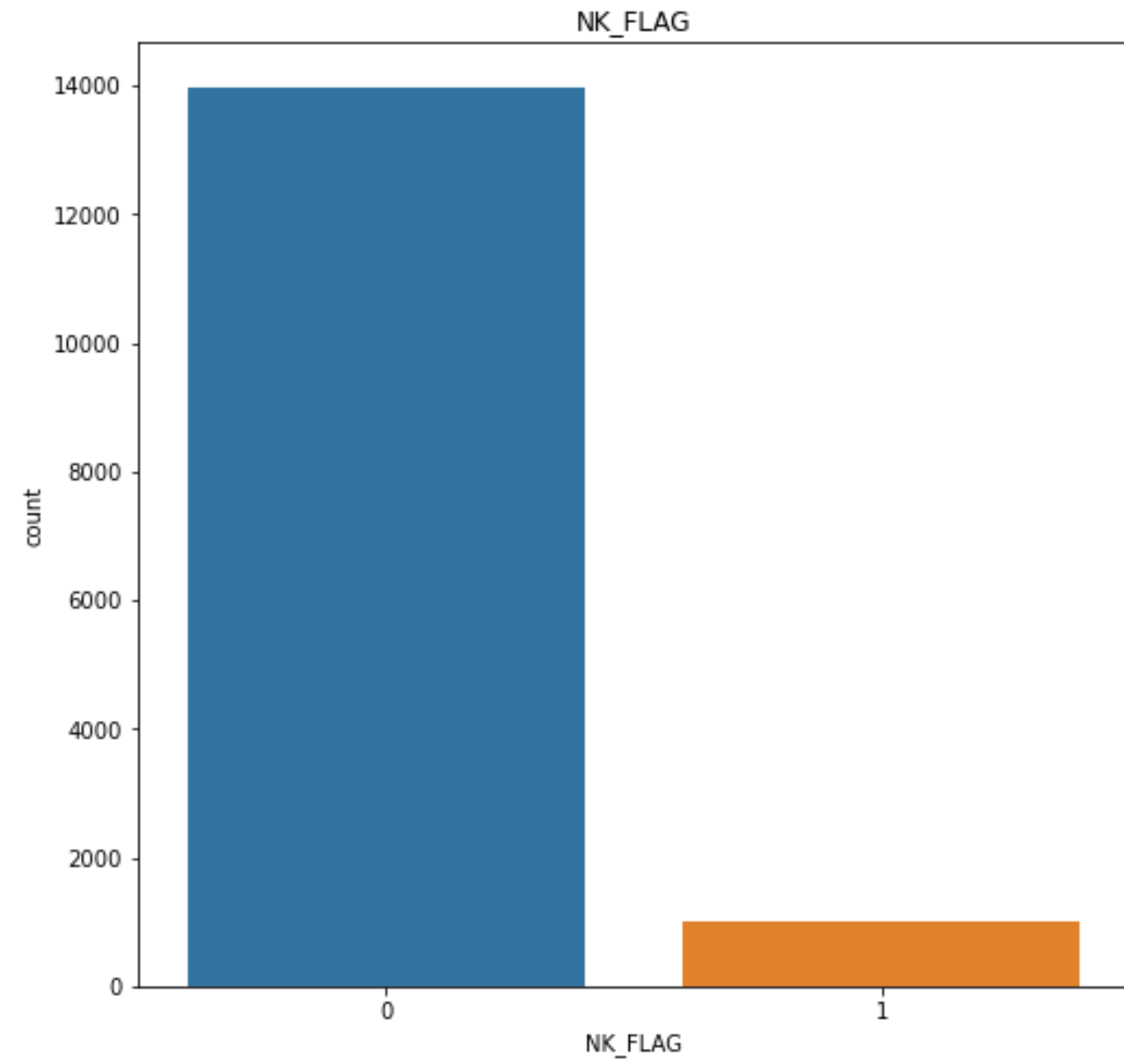
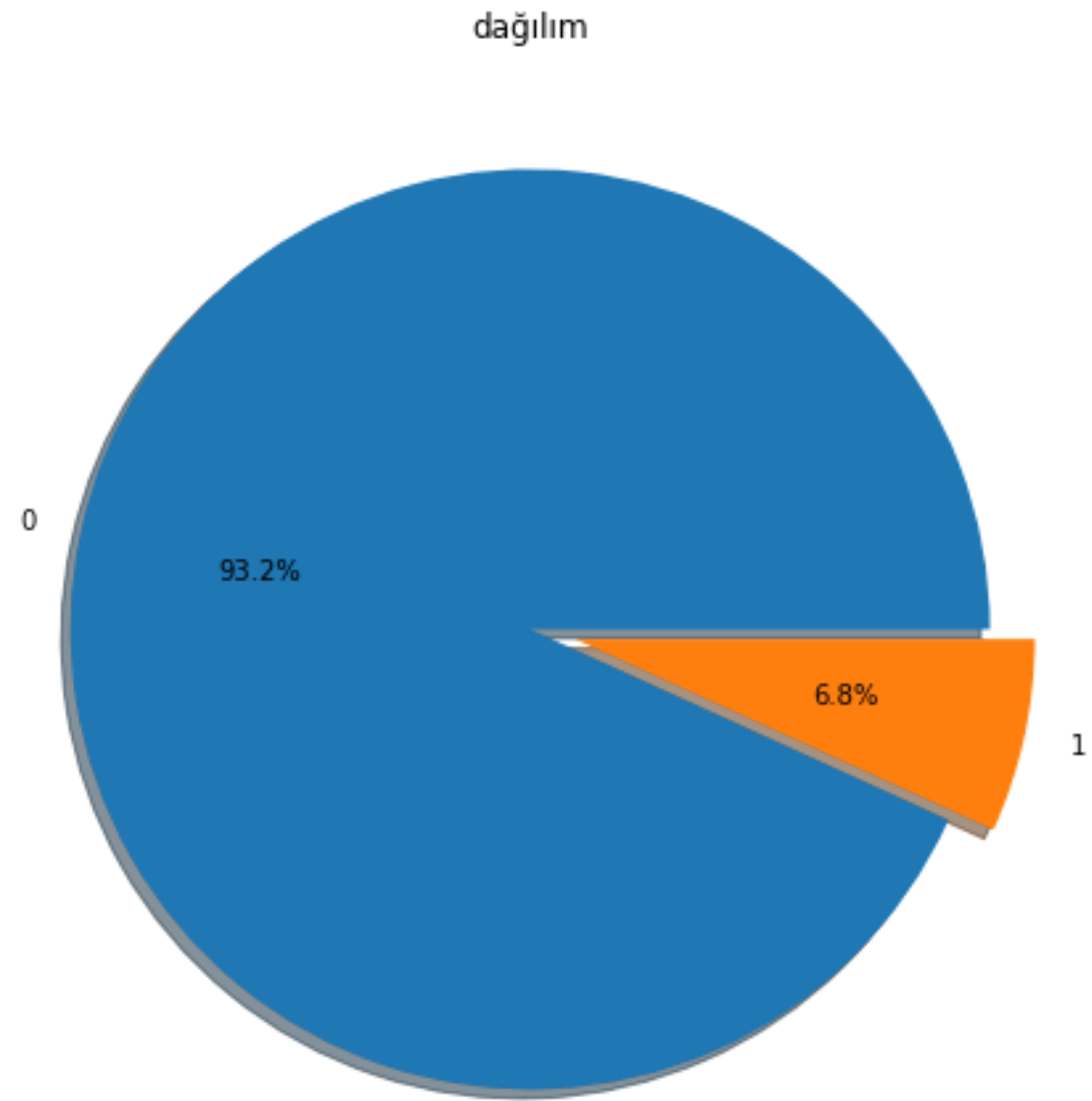


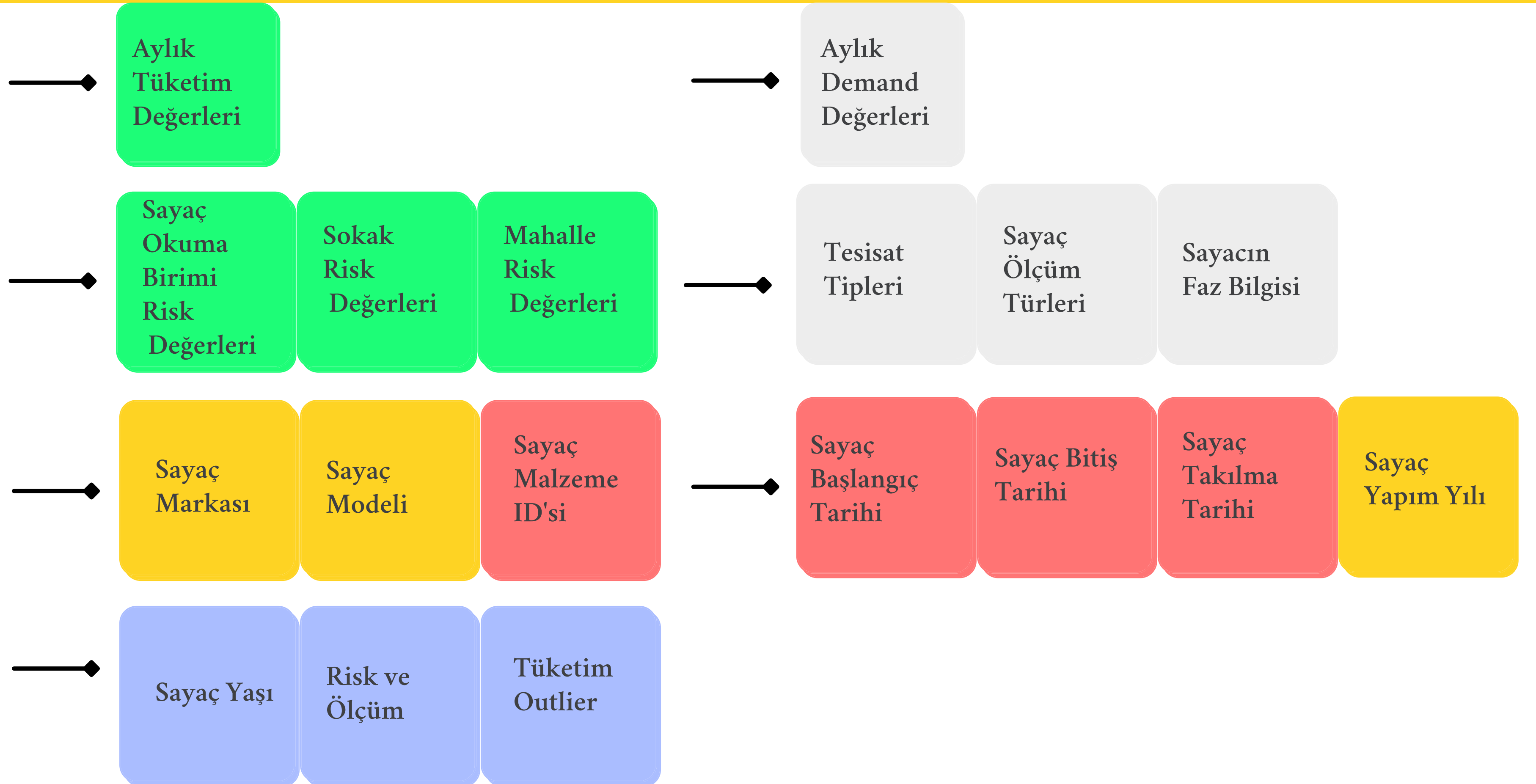
Modelleme

4

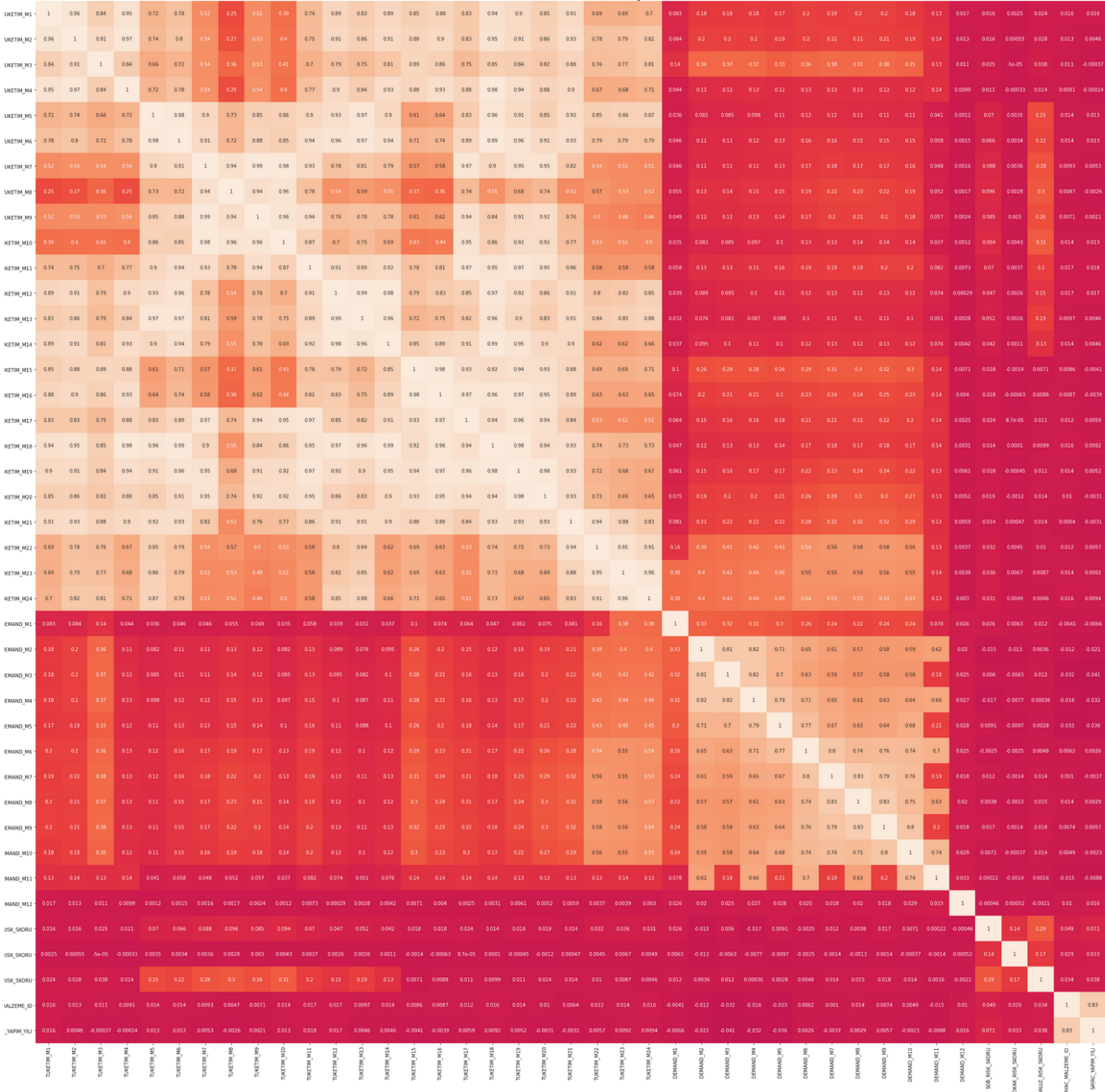


Skor
İyileştirme



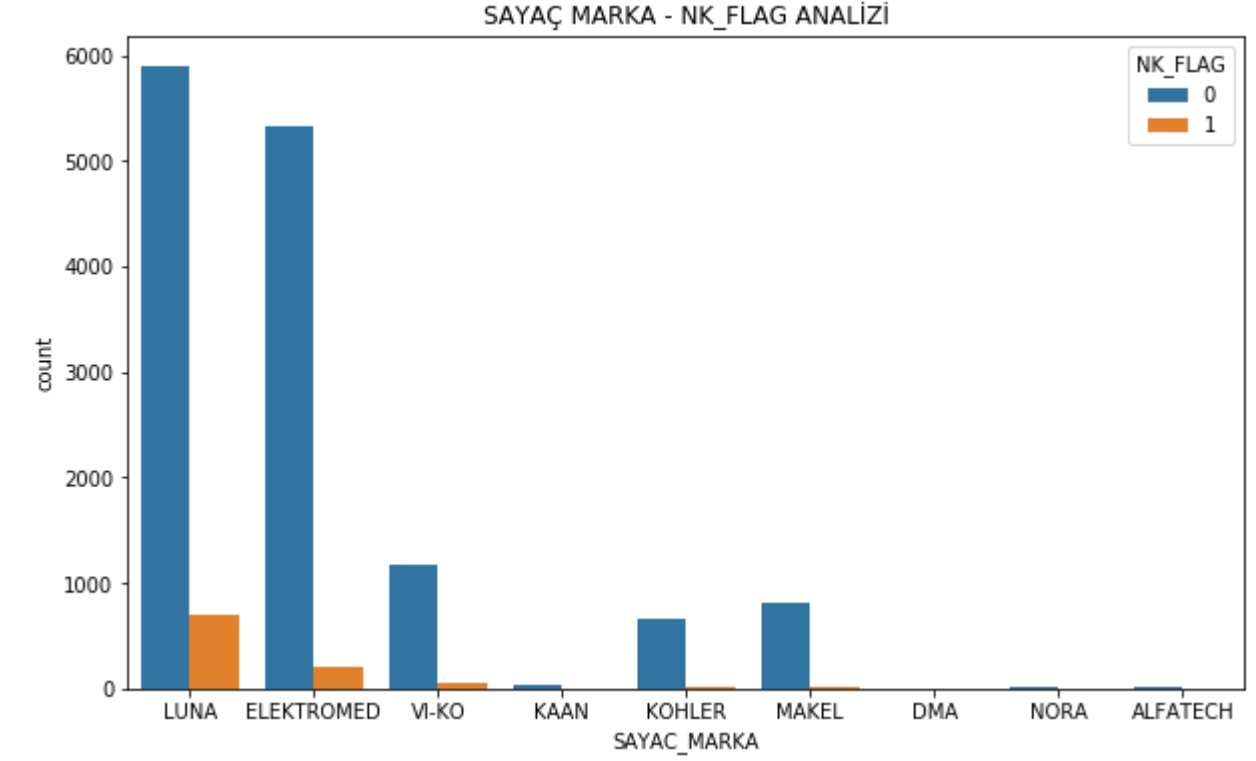
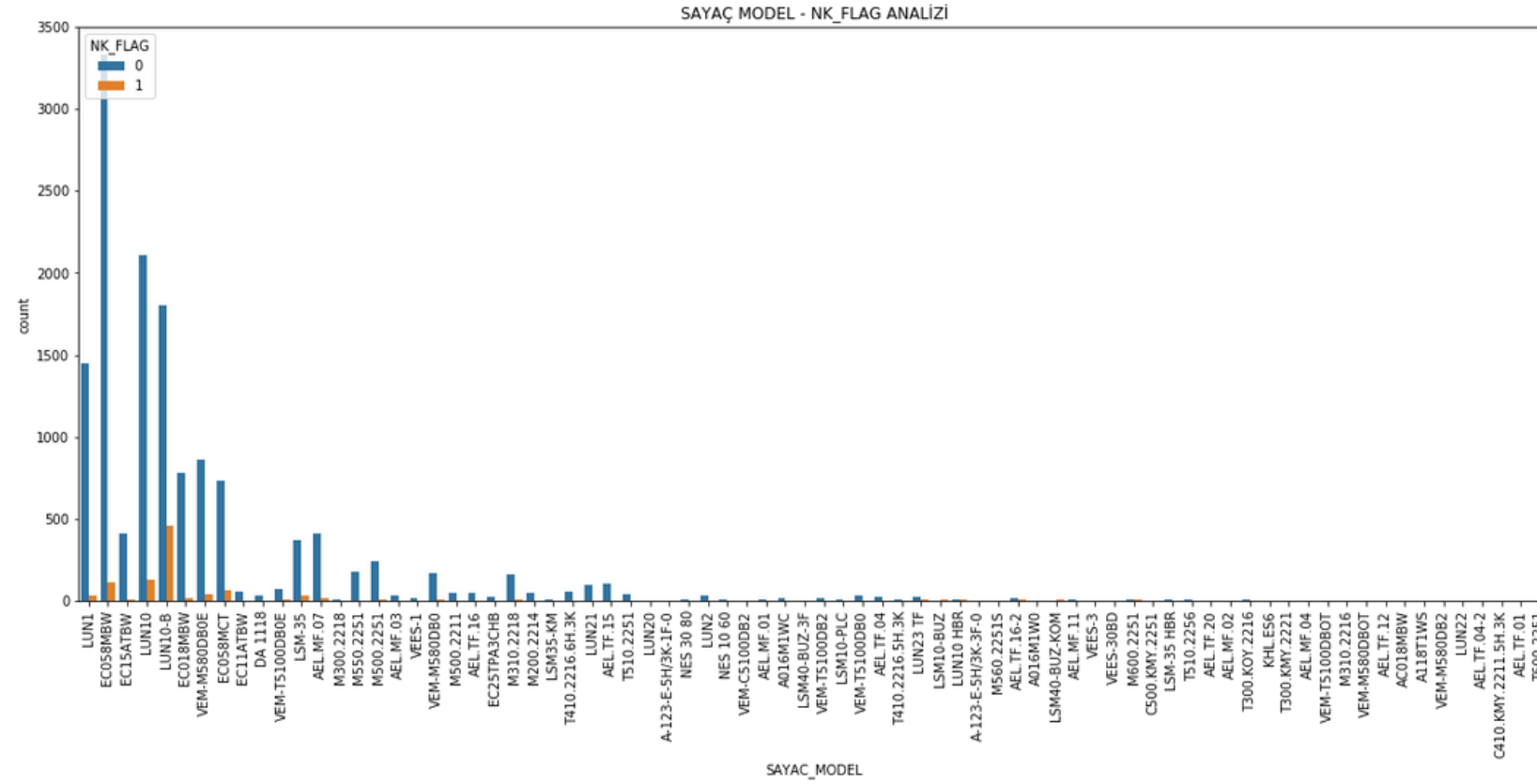
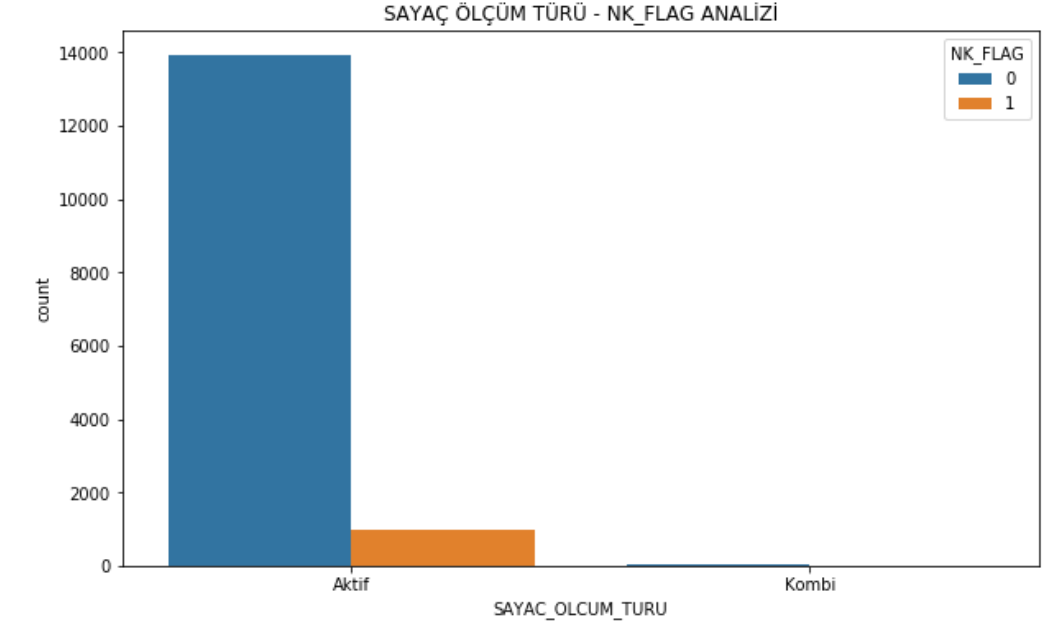
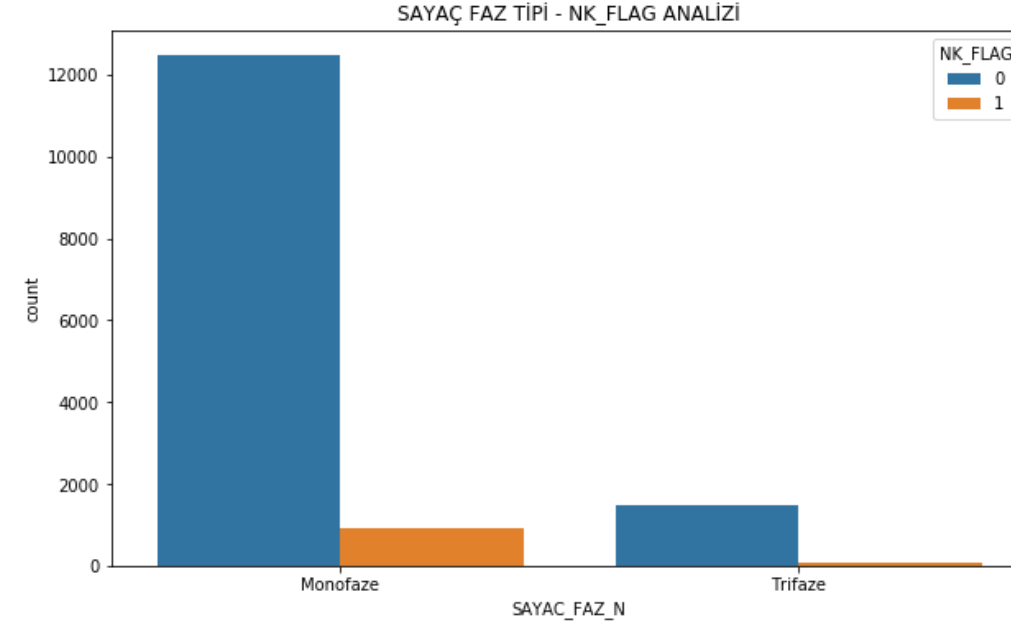
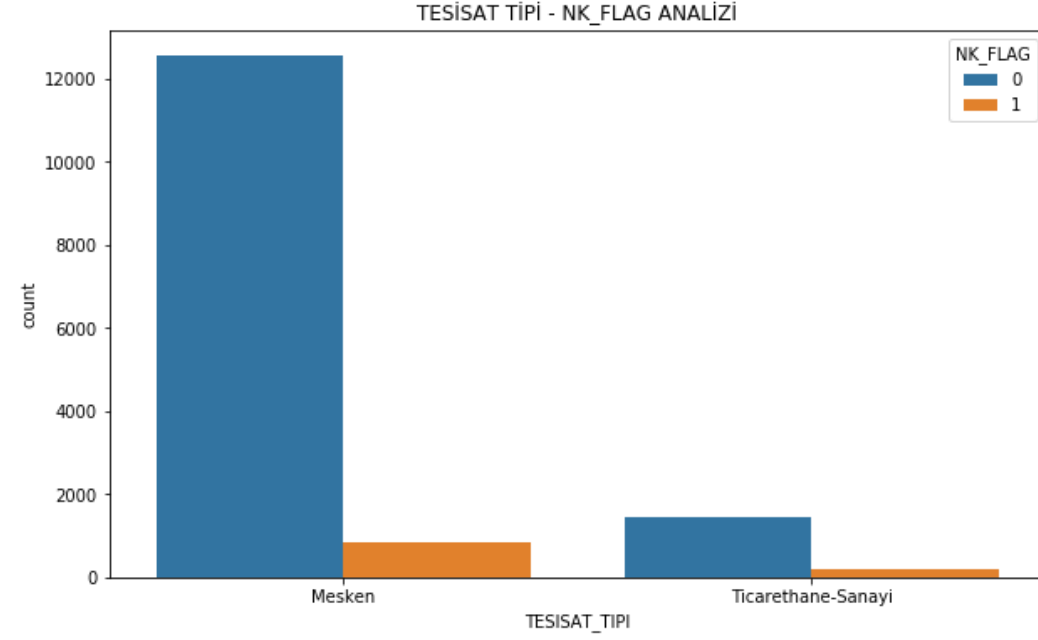


Correlation Heatmap



Correlation Heatmap

TUKETIM_M1	1	0.96	0.84	0.95	0.72	0.78	0.52	0.25	0.52	0.39	0.74	0.89	0.83	0.89	0.85	0.88	0.83	0.94	0.9	0.85	0.91	0.69	0.69	0.7
TUKETIM_M2	0.96	1	0.91	0.97	0.74	0.8	0.54	0.27	0.53	0.4	0.75	0.91	0.86	0.91	0.88	0.9	0.83	0.95	0.91	0.86	0.93	0.78	0.79	0.82
TUKETIM_M3	0.84	0.91	1	0.84	0.66	0.72	0.54	0.36	0.53	0.41	0.7	0.79	0.75	0.81	0.89	0.86	0.75	0.85	0.84	0.82	0.88	0.76	0.77	0.81
TUKETIM_M4	0.95	0.97	0.84	1	0.72	0.78	0.54	0.25	0.54	0.4	0.77	0.9	0.84	0.93	0.88	0.93	0.88	0.98	0.94	0.88	0.9	0.67	0.68	0.71
TUKETIM_M5	0.72	0.74	0.66	0.72	1	0.98	0.9	0.73	0.85	0.86	0.9	0.93	0.97	0.9	0.61	0.64	0.83	0.96	0.91	0.85	0.92	0.85	0.86	0.87
TUKETIM_M6	0.78	0.8	0.72	0.78	0.98	1	0.91	0.72	0.88	0.85	0.94	0.96	0.97	0.94	0.71	0.74	0.89	0.99	0.96	0.91	0.93	0.79	0.79	0.79
TUKETIM_M7	0.52	0.54	0.54	0.54	0.9	0.91	1	0.94	0.99	0.98	0.93	0.78	0.81	0.79	0.57	0.58	0.97	0.9	0.95	0.95	0.82	0.54	0.52	0.51
TUKETIM_M8	0.25	0.27	0.36	0.25	0.73	0.72	0.94	1	0.94	0.96	0.78	0.54	0.59	0.55	0.37	0.36	0.74	0.55	0.68	0.74	0.52	0.57	0.53	0.52
TUKETIM_M9	0.52	0.53	0.53	0.54	0.85	0.88	0.99	0.94	1	0.96	0.94	0.76	0.78	0.78	0.61	0.62	0.94	0.84	0.91	0.92	0.76	0.5	0.48	0.46
TUKETIM_M10	0.39	0.4	0.41	0.4	0.86	0.85	0.98	0.96	0.96	1	0.87	0.7	0.75	0.69	0.43	0.44	0.95	0.86	0.92	0.92	0.77	0.53	0.51	0.5
TUKETIM_M11	0.74	0.75	0.7	0.77	0.9	0.94	0.93	0.78	0.94	0.87	1	0.91	0.89	0.92	0.78	0.81	0.97	0.95	0.97	0.95	0.86	0.58	0.58	0.58
TUKETIM_M12	0.89	0.91	0.79	0.9	0.93	0.96	0.78	0.54	0.76	0.7	0.91	1	0.99	0.98	0.79	0.83	0.85	0.97	0.92	0.86	0.91	0.8	0.82	0.85
TUKETIM_M13	0.83	0.86	0.75	0.84	0.97	0.97	0.81	0.59	0.78	0.75	0.89	0.99	1	0.96	0.72	0.75	0.82	0.96	0.9	0.83	0.91	0.84	0.85	0.88
TUKETIM_M14	0.89	0.91	0.81	0.93	0.9	0.94	0.79	0.55	0.78	0.69	0.92	0.98	0.96	1	0.85	0.89	0.91	0.99	0.95	0.9	0.9	0.62	0.62	0.66
TUKETIM_M15	0.85	0.88	0.89	0.88	0.61	0.71	0.57	0.37	0.61	0.43	0.78	0.79	0.72	0.85	1	0.98	0.93	0.92	0.94	0.93	0.88	0.69	0.69	0.71
TUKETIM_M16	0.88	0.9	0.86	0.93	0.64	0.74	0.58	0.36	0.62	0.44	0.81	0.83	0.75	0.89	0.98	1	0.97	0.96	0.97	0.95	0.89	0.63	0.63	0.65
TUKETIM_M17	0.83	0.83	0.75	0.88	0.83	0.89	0.97	0.74	0.94	0.95	0.97	0.85	0.82	0.91	0.93	0.97	1	0.94	0.96	0.94	0.84	0.53	0.52	0.51
TUKETIM_M18	0.94	0.95	0.85	0.98	0.96	0.99	0.9	0.55	0.84	0.86	0.95	0.97	0.96	0.99	0.92	0.96	0.94	1	0.98	0.94	0.93	0.74	0.73	0.73
TUKETIM_M19	0.9	0.91	0.84	0.94	0.91	0.96	0.95	0.68	0.91	0.92	0.97	0.92	0.9	0.95	0.94	0.97	0.96	0.98	1	0.98	0.93	0.72	0.68	0.67
TUKETIM_M20	0.85	0.86	0.82	0.88	0.85	0.91	0.95	0.74	0.92	0.92	0.95	0.86	0.83	0.9	0.93	0.95	0.94	0.94	0.98	1	0.93	0.73	0.69	0.65
TUKETIM_M21	0.91	0.93	0.88	0.9	0.92	0.93	0.82	0.52	0.76	0.77	0.86	0.91	0.91	0.9	0.88	0.89	0.84	0.93	0.93	0.93	1	0.94	0.88	0.83
TUKETIM_M22	0.69	0.78	0.76	0.67	0.85	0.79	0.54	0.57	0.5	0.53	0.58	0.8	0.84	0.62	0.69	0.63	0.53	0.74	0.72	0.73	0.94	1	0.95	0.91
TUKETIM_M23	0.69	0.79	0.77	0.68	0.86	0.79	0.52	0.53	0.48	0.51	0.58	0.82	0.85	0.62	0.69	0.63	0.52	0.73	0.68	0.69	0.88	0.95	1	0.96
TUKETIM_M24	0.7	0.82	0.81	0.71	0.87	0.79	0.51	0.52	0.46	0.5	0.58	0.85	0.88	0.66	0.71	0.65	0.51	0.73	0.67	0.65	0.83	0.91	0.96	1



ANALİZ SORULARI

- 1 Tüketim verilerinde tüm satırın sıfır olması ne ifade ediyor?
- 2 Demand ve tüketim verileri arasında bir ilişki var mı?
- 3 Sokak risk skorunun mahalle risk skorundan büyük olması bizi kaçağa götürür mü?
- 4 Ticarethanede daha çok trifaze görülürken, meskende monofaze çoğunlukta görülmesi eksik veri doldurmada işimize yarar mı?
Tüketim değerleri için 20000 eşik noktasından sonraki değerler 1'i yakalıyor. Burdan yeni bir kolon mu oluşturulmalı yoksa yeni gelen veride bu aralıkta olan değerler predict edilmeden sonuç mu üretilmeli?
- 5

KULLANDIĞIMIZ MODELLER

- SVC
- Cat Boost
- Random Forest
- Decision Tree

MODEL DEZANATAJLARI

RANDOM FORESTS

- Random Forests, birden fazla karar ağacına sahip olduğu için tahmin üretmede yavaştır. Ne zaman bir tahminde bulunursa, ormandaki tüm ağaçların verilen aynı girdi için bir tahmin yapması ve ardından oylama yapması gerekir. Tüm bu süreç zaman alıcıdır.
- Ağaçtaki yolu izleyerek kolayca karar verebileceğiniz bir karar ağacına kıyasla modeli yorumlamak zordur.

DECISION TREES

- Sürekli nitelik değerlerini tahmin etmekte çok başarılı değildir
- Sınıf sayısı fazla ve öğrenme kümesi örnekleri sayısı az olduğunda model oluşturma çok başarılı değildir
- Zaman ve yer karmaşıklığı öğrenme kümesi örnekleri sayısına, nitelik sayısına ve oluşan ağacın yapısına bağlıdır
- Hem ağaç oluşturma karmaşıklığı hem de ağaç budama karmaşıklığı fazladır

SVC

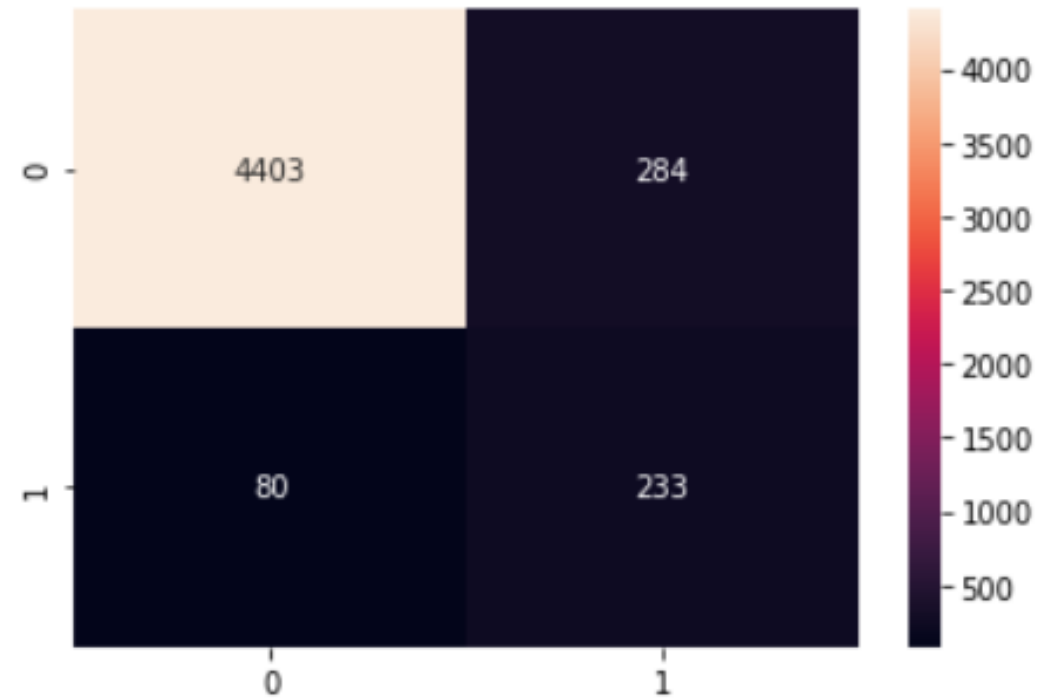
- Bir çok kernel fonksiyonunu kullanabilmemize imkan sağlıyor
- Büyük datasetler için uzun training süresi gerektiriyor.
- Nihai modelin çıktıları ve future importance kurgularının modele nasıl yansıdığı kolay yorumlanabilir değil. Bu nedenlerden dolayı, hyperparameter yönelik hassas ayarlamaları gerçekleştirmek zorlaşmakta

MODEL SKOR TABLOSU

Model İsmi	Accuracy	Precision		Recall		F1-Score	
		0	1	0	1	0	1
SVC w/0.34 probability threshold, kernel = linear	84%	99%	29%	83%	92%	91%	44%
SVC kernel = linear	87%	99%	34%	87%	90%	93%	49%
Decision Tree Classifier	93%	98%	51%	95%	72%	96%	60%
Decision Tree Classifier w/0.4 probability threshold	94%	98%	53%	95%	73%	96%	61%
SVC kernel = rbf	93%	99%	48%	93%	87%	96%	62%
Random Forest w/0.34 probability threshold	94%	99%	56%	95%	83%	97%	67%
SVC w/0.81 probability threshold, kernel = rbf	96%	98%	67%	97%	80%	98%	72%
Random Forest Classifier	97%	98%	81%	99%	71%	98%	76%
CatBoost Classifier (max_depth = 10, n_estimators = 100)	96%	99%	67%	97%	88%	98%	76%
CatBoost Classifier (max_depth = 5, n_estimators = 100)	96%	99%	71%	97%	83%	98%	77%
CatBoost Classifier (max_depth = 5, n_estimators = 200)	96%	99%	70%	97%	88%	98%	78%
CatBoost Classifier w/0.75 probability threshold (max_depth = 5, n_estimators = 200)	97%	99%	83%	99%	81%	99%	82%
CatBoost Classifier w/0.75 probability threshold (max_depth = 5, n_estimators = 300)	98%	99%	86%	99%	81%	99%	84%

MODEL SONUCU

	precision	recall	f1-score	support
0	0.98	0.94	0.96	4687
1	0.45	0.74	0.56	313
accuracy			0.93	5000
macro avg	0.72	0.84	0.76	5000
weighted avg	0.95	0.93	0.94	5000



Kullandığımız modeller arasında en yüksek skoru veren Cat Boost Algoritması olduğu için bu modeli kullanmaya karar verdik.

EKİBİMİZ




İrfan
Türkmen


 /irfanturkmen/

 irfan.turkmen@outlook.com



Selin
Ünlü

 /selin-unlu/

 slinunlu@gmail.com




Kübra
Yiğiter


 /kubrayigiter/

 kubrayigiter01@gmail.com




Kemal Burak
Arıboğa


 /kemalburakariboga/

 kemalburak94@gmail.com



Eylül
Akkurt

 /eylül-akkurt8/

 akkurteylul98@outlook.com