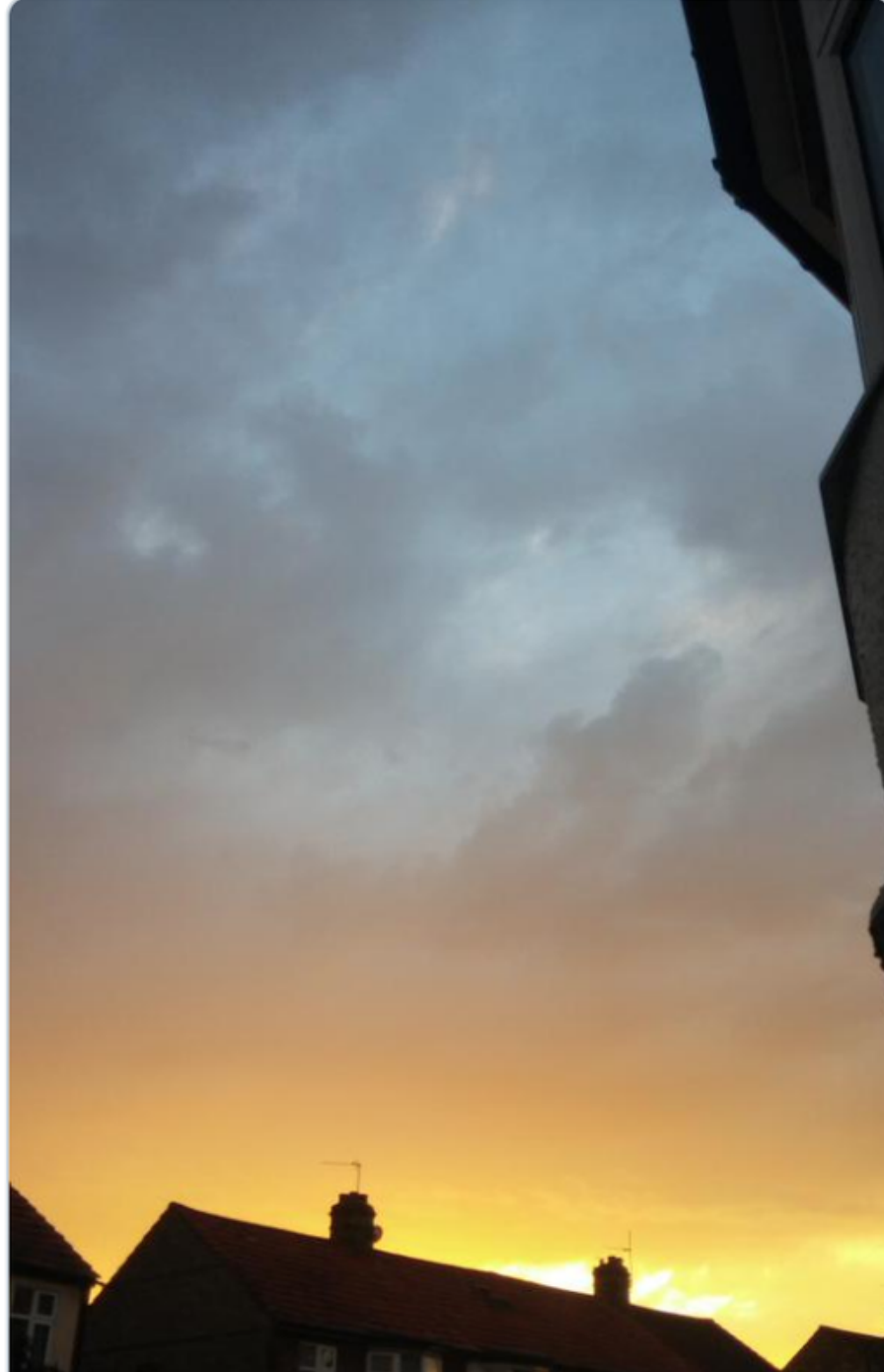


NATURAL LANGUAGE PROCESSING WITH DISASTER TWEETS

Sunan: Sercan Sözen



On plus side LOOK AT THE SKY LAST NIGHT IT WAS ABLAZE



Proje Konusu

Twitter, acil durumlarda önemli bir iletişim kanalı haline geldi. Akıllı telefonların her yerde bulunması, insanların gözlemledikleri bir acil durumu gerçek zamanlı olarak duyurmalarını sağlıyor.

Ancak, bir kişinin sözlerinin gerçekten bir felaketi tarif edip etmediği her zaman net değildir.

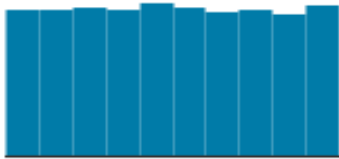

Yazar, “ABLAZE” kelimesini açıkça kullanmakta ancak bunu mecazi anlamda ifade etmektedir. Bu, özellikle görselin yardımıyla, bir insan için açıktır. Ancak bir makine için karmaşıktır.

Amaç

Bu yarışmada, hangi Tweetlerin gerçek felaketlerle ilgili olduğunu ve hangilerinin olmadığını tahmin eden bir makine öğrenimi modeli oluşturmamız isteniyor. Elle sınıflandırılmış 10.000 tweet'lik bir veri kümesi mevcut.



Veri Seti

🔍 id	📄 keyword	📄 location	📄 text	# target
 110.9k	222 unique values	[null] 33% USA 1% Other (4976) 65%	7503 unique values	 01
144	accident	UK	.@NorwayMFA #Bahrain police had previously died in a road accident they were not killed by explosion...	1
145	accident	Nairobi, Kenya	I still have not heard Church Leaders of Kenya coming forward to comment on the accident issue and d...	0
146	aftershock	Instagram - @heyimginog	@afterShock_DeLo scuf ps live and the game... cya	0

Train Veri Seti

🔍 id	📄 keyword	📄 location	📄 text
 3262.50 - 4350.00 Count: 333 010.9k	222 unique values	[null] 34% New York 1% Other (2120) 65%	3243 unique values
111	accident	Bexhill	@Traffic_SouthE @roadpol_east Accident on A27 near Lewes is it Kingston Roundabout rather than A283
115	accident	Anime World	@sakuma_en If you pretend to feel a certain way the feeling can become genuine all by accident. -Hei...
116	accident		For Legal and Medical Referral Service @1800_Injured Call us at: 1-800-465-87332 #accident #slipandf...

Test Veri Seti

Veri Seti

Test veri seti etiketli bir veri seti olmadığından supervised learning yapmak için train veri seti tekrar train, validation ve test (.7, .15 ve .15) olacak şekilde bölümlendi. Bu bölümleme sklearn kütüphanesinin train_test_split metodu 2 kez kullanılarak elde edildi. Kullanılan train veri setinde 4307 negatif, 3196 pozitif olmak üzere toplam 7503 adet veri bulunuyor.

[illegible]

Verinin Önişlenmesi

NLTK (natural language tool kit) ve string benzeri kütüphaneler kullanılarak veri setleri; noktalama işaretleri, stopwordler ve hmtl etiketleri gibi gereksiz kelimelerden ve boş satırlardan ayıklandı.

	id	keyword	location	text	target
0	1	NaN	NaN	Our Deeds are the Reason of this #earthquake M...	1
1	4	NaN	NaN	Forest fire near La Ronge Sask. Canada	1
2	5	NaN	NaN	All residents asked to 'shelter in place' are ...	1
3	6	NaN	NaN	13,000 people receive #wildfires evacuation or...	1
4	7	NaN	NaN	Just got sent this photo from Ruby #Alaska as ...	1
5	8	NaN	NaN	#RockyFire Update => California Hwy. 20 closed...	1
6	10	NaN	NaN	#flood #disaster Heavy rain causes flash flood...	1
7	13	NaN	NaN	I'm on top of the hill and I can see a fire in...	1
8	14	NaN	NaN	There's an emergency evacuation happening now ...	1
9	15	NaN	NaN	I'm afraid that the tornado is coming to our a...	1

Ayıklanmamış Veri

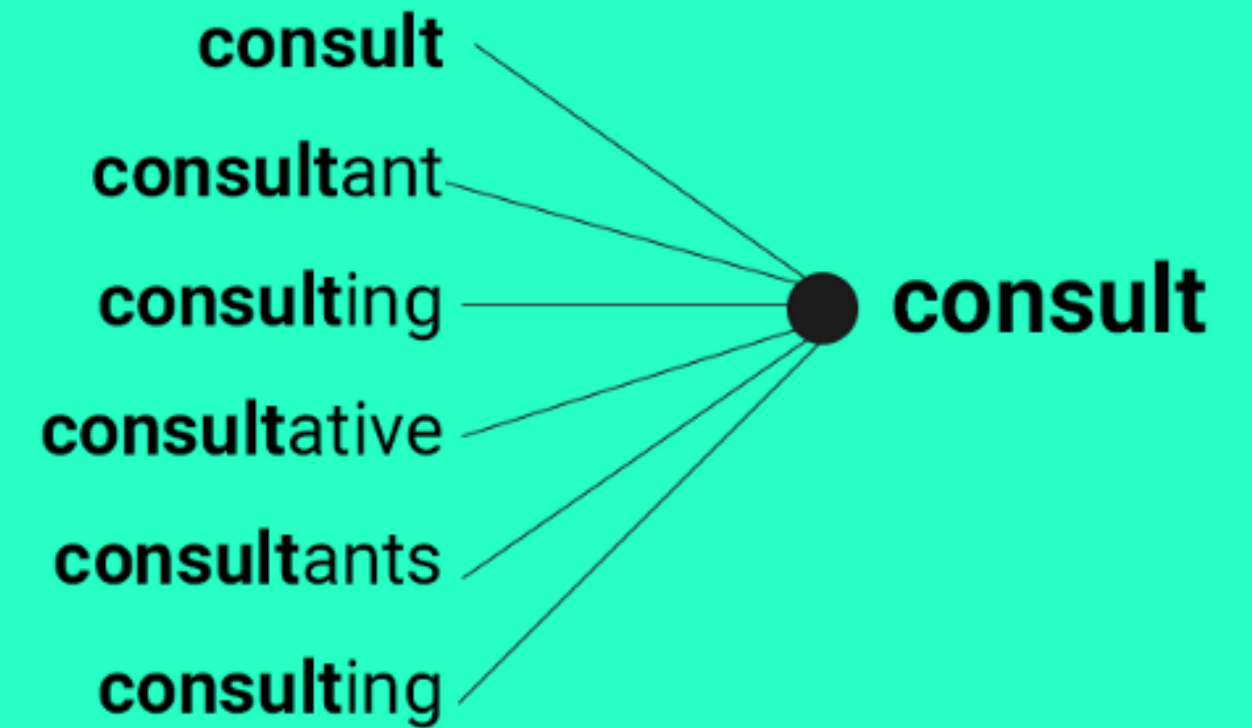
	id	keyword	location	text	target
0	1	NaN	NaN	deed reason earthquak may allah forgiv u	1
1	4	NaN	NaN	forest fire near la rong sask canada	1
2	5	NaN	NaN	resid ask shelter place notifi offic evacu she...	1
3	6	NaN	NaN	peopl receiv wildfir evacu order california	1
4	7	NaN	NaN	got sent photo rubi alaska smoke wildfir pour ...	1
5	8	NaN	NaN	rockyfir updat california hwi close direct due...	1
6	10	NaN	NaN	flood disast heavi rain caus flash flood stree...	1
7	13	NaN	NaN	top hill see fire wood	1
8	14	NaN	NaN	emerg evacu happen build across street	1
9	15	NaN	NaN	afraid tornado come area	1

Ayıklanmış Veri

Yöntem

Ayıklanan veriden oluşan dataframe, Countvectorizer (bag of words) ve TF-IDF vectorizer fonksiyonları ile vektöre dönüştürüldü. Maksimum özellik (kelime) sayısı 4000 olarak belirlendi. Uygulanacak modeller bu iki vektörizasyon yöntemiyle de test edildi. Başarıları kıyaslandı.

Buna ek olarak aynı veri seti K-Fold Cross Validation yöntemiyle K=5 ve K=10 katlama için sınıflandırıcıların başarıları test edildi.



Kullanılan Sınıflandırıcılar

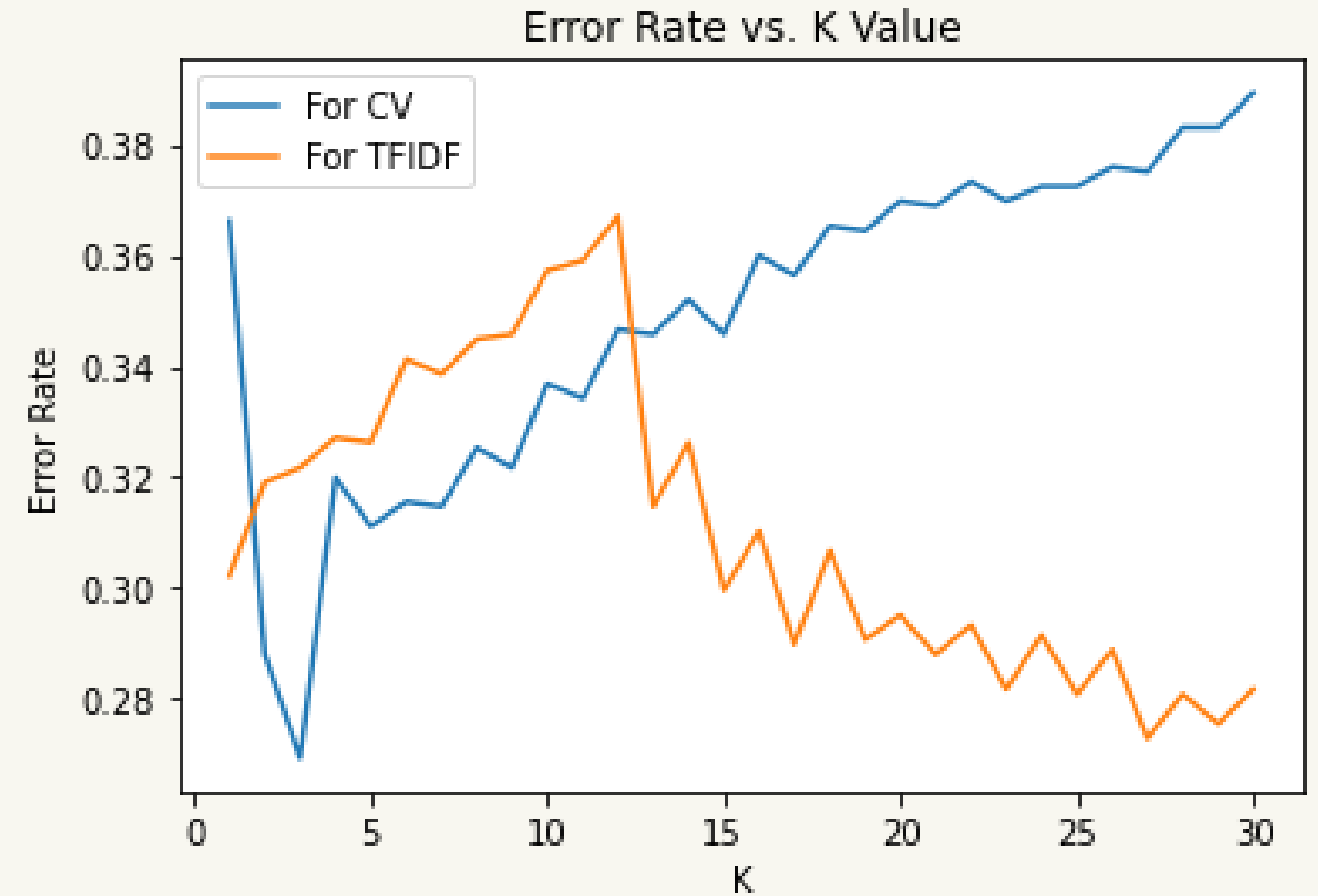
- Naive Bayes
 - Gaussian NB ve Multinomial NB
- Decision Tree
- K-Nearest Neighbor
- Logistic Regression
- Multilayer Perceptron



Parametre Ayarlamaları

Her bir sınıflandırıcı için validasyon seti üzerinde tahminlemeler yapılarak başarı oranları ölçüldü. Daha sonra gerekli yerlerde sınıflandırıcının parametleri düzenlenerek daha iyi bir sonuç elde edilmeye çalışıldı. Örneğin K en yakın komşu sınıflandırıcı için 1-31 arasında komşu sayısı değerleri denendi.

K en yakın komşu sınıflandırıcısının,
Bag of Words ile vektörize edilmiş veri üzerinde en az hata oranını 3 komşulukta
TF - IDF ile vektörize edilmiş veri üzerinde ise en az hata oranını 27 komşulukta yakaladığını gözlemliyoruz.



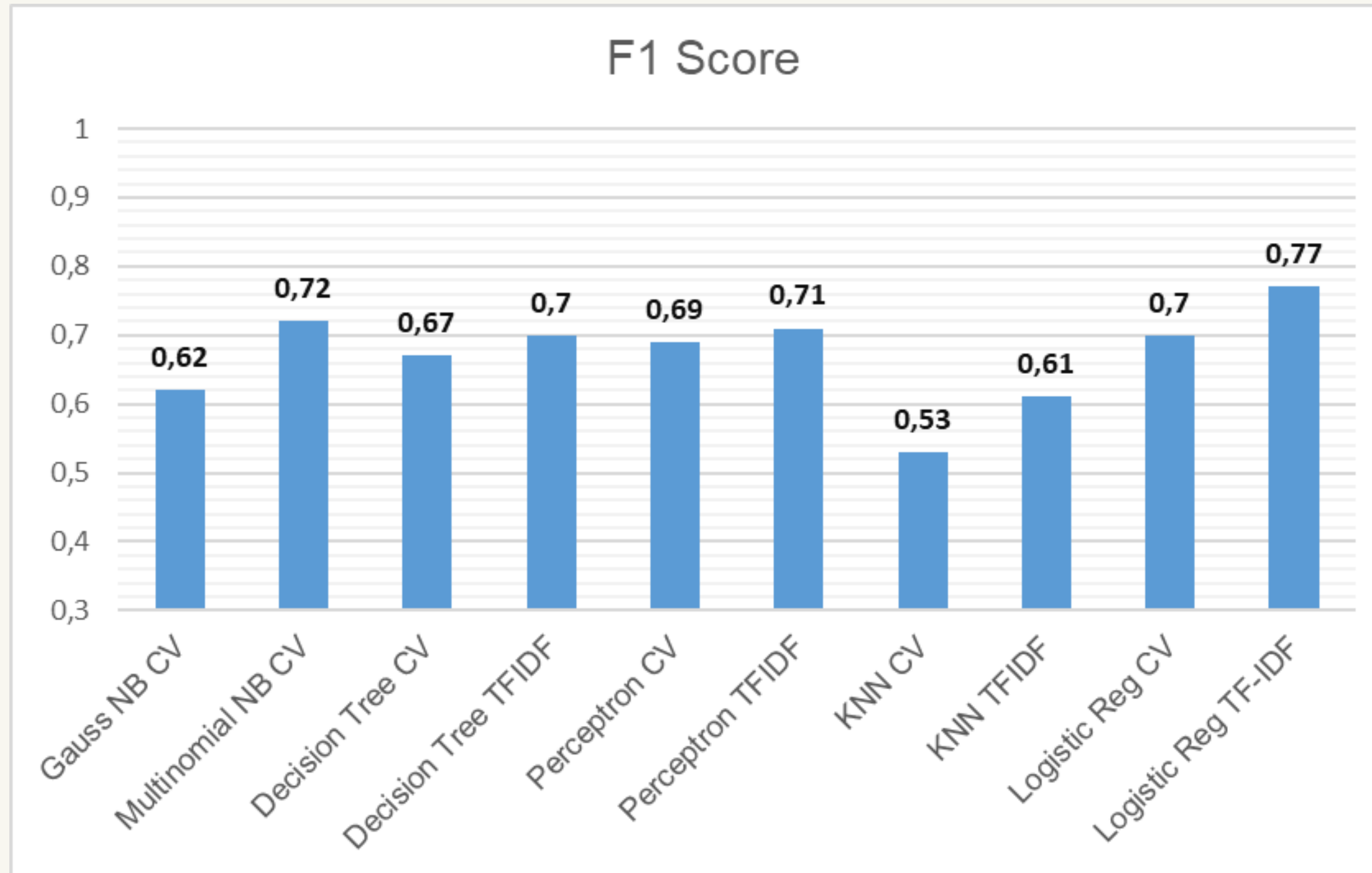
Parametre Ayarlamaları

Karar ağacı için maksimum ağaç derinliği 70, maksimum yaprak düğüm sayısı 100 olarak belirlendi.

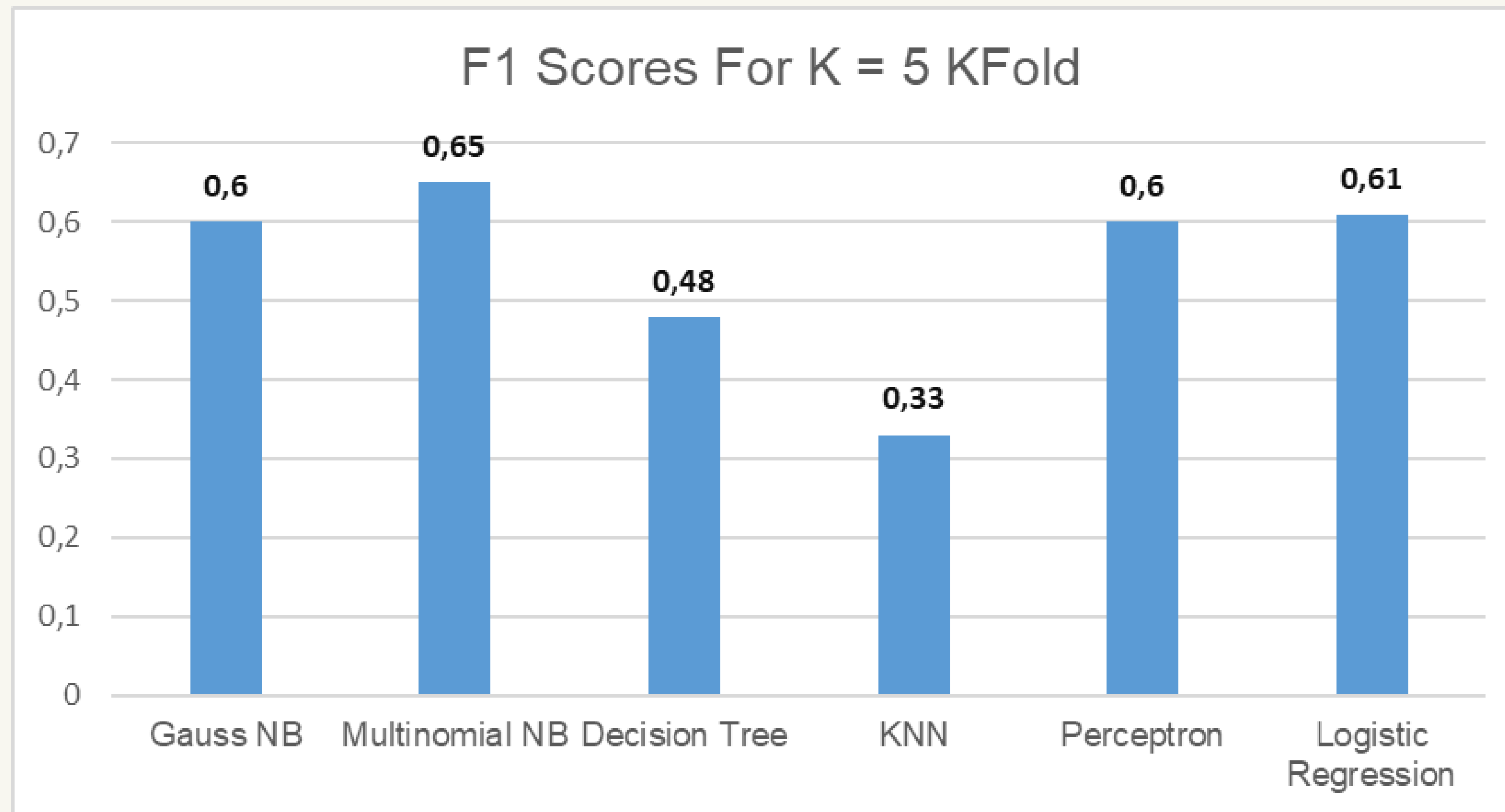
Çok katmanlı perceptron için aktivasyon fonksiyonu relu, çözücü olarak sigmoid fonksiyonu, başlangıç learning rate olarak 0.1 değeri belirlendi. Ayrıca learning rate adaptive olarak ayarlandı. Kullanılan 2 adet gizli katmandaki nöron sayıları da sırasıyla 5 ve 10 olarak belirlendi.

Lojistik regresyon fonksiyonunda sadece sınıf ağırlığı parametresi "balanced" olarak belirlendi.

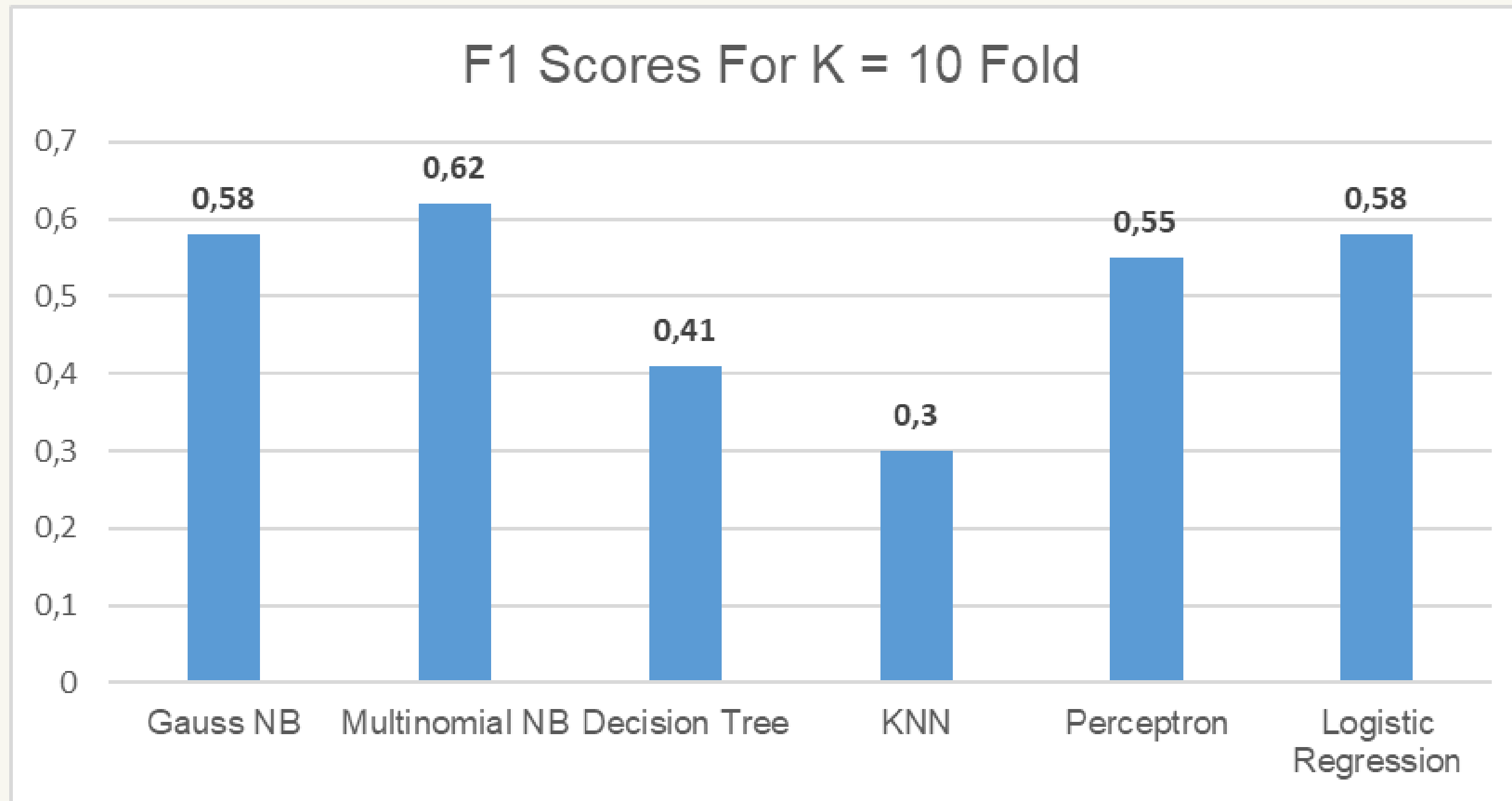
Deneyler



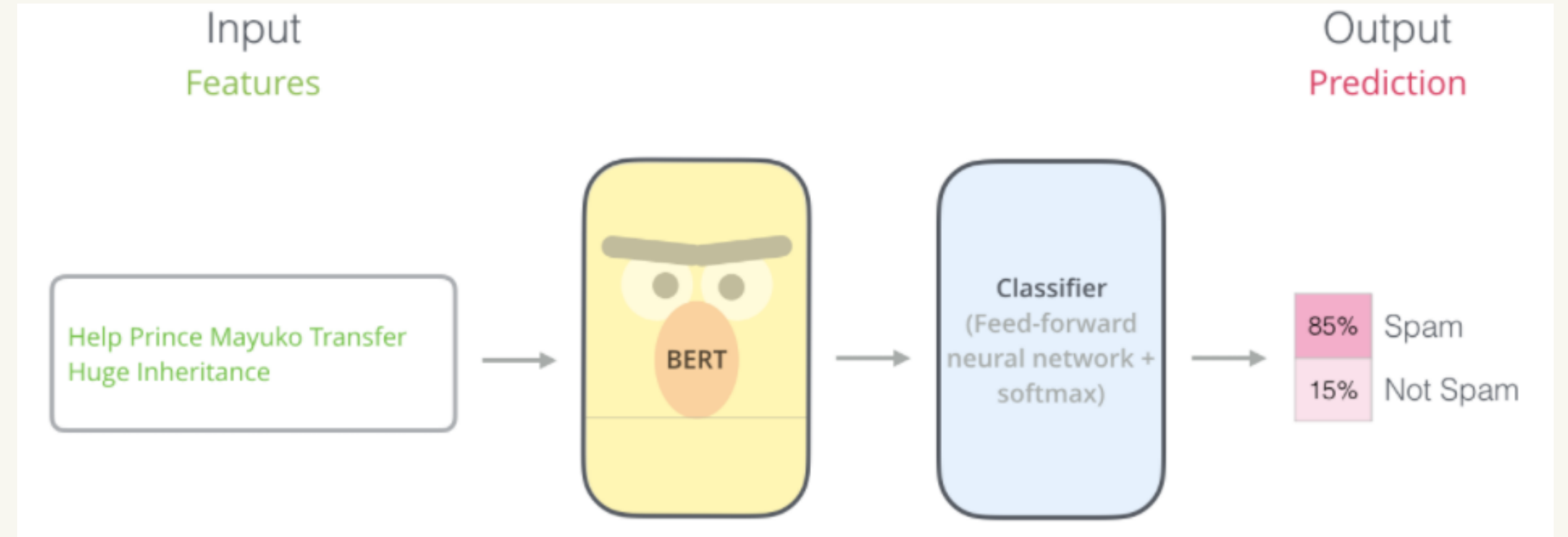
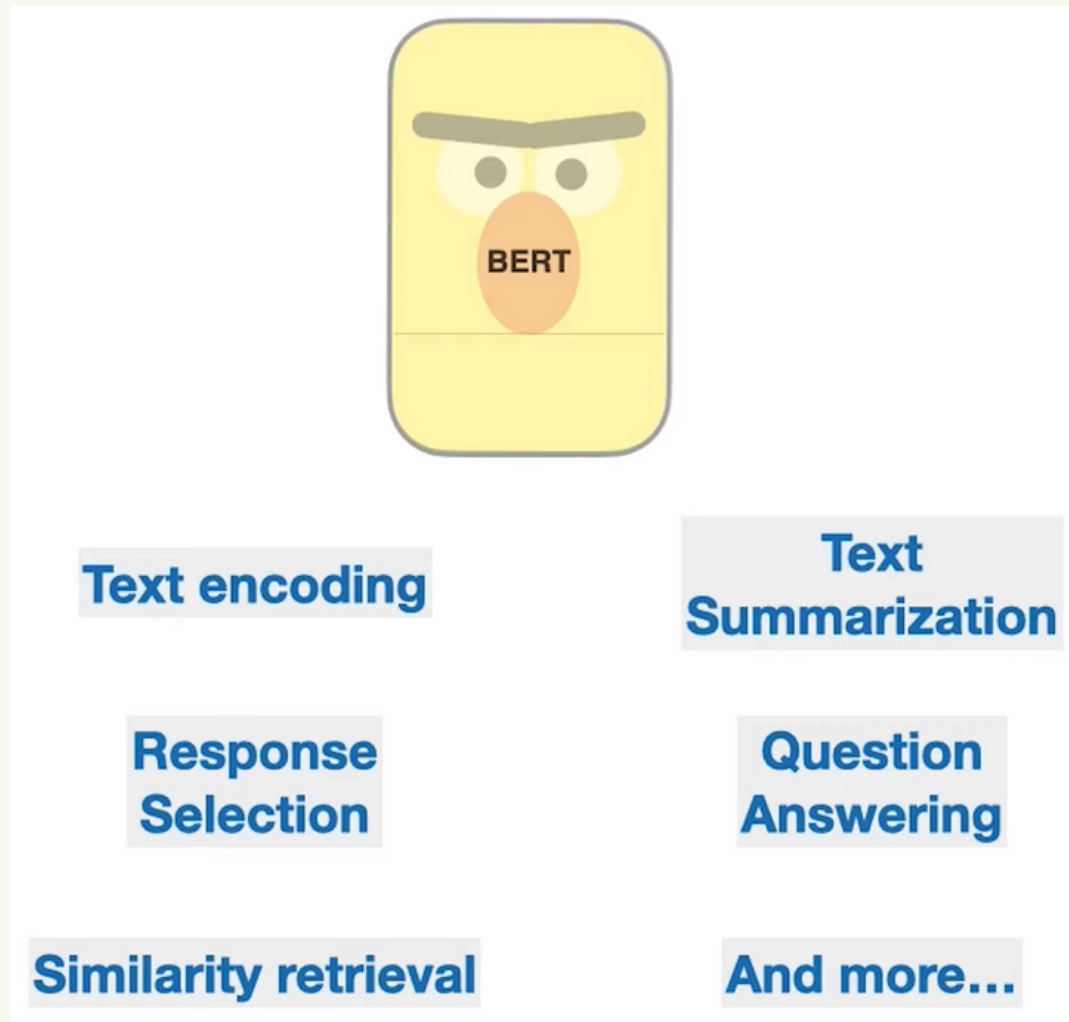
Deneyler



Deneyler



BERT (Bidirectional Encoder Representations from Transformers)



6 encoder katmanı, 768 hidden unit
2.5 B Wikipedia
800 M BookCorpus verisiyle
4 adet TPU ile 4 gün eğitilmiş

BERT (Bidirectional Encoder Representations from Transformers)

