MAKİNE ÖĞRENMESİ İLE SPAM SINIFLANDIRMASI

GİRİŞ

Günümüz teknolojisinin geldiği seviye dolayısıyla, dijital kaynaklar hayatımızda oldukça büyük bir yer edinmeye başlamış ve neredeyse bir uzuv haline gelmiştir. Bu kaynakların iletişim alanındaki kullanımı oldukça ileri bir seviyeye gelmiş olup amacı dışındaki kullanımı pek çok maddi yahut manevi zararlar içerebilecek olumsuz durumu beraberinde getirmiştir. Bu çalışmada işlenecek olan spam mesajları da bu amaç dışı kullanıma bir örnektir.

AMAÇ

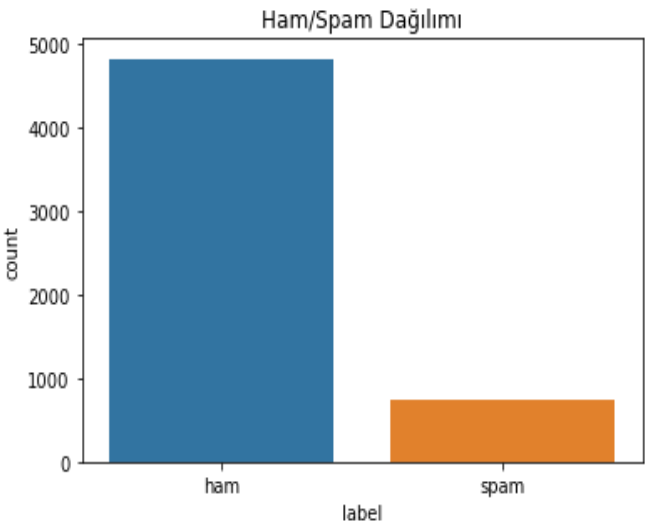
Proje kapsamında, kullanıcının gereksiz, uygunsuz yahut kötü niyetli mesajlar dolayısıyla herhangi bir zarara uğramaması amaçlanmış olup bu doğrultuda makine öğrenmesi destekli bir spam filtreleme uygulaması geliştirilmiştir.

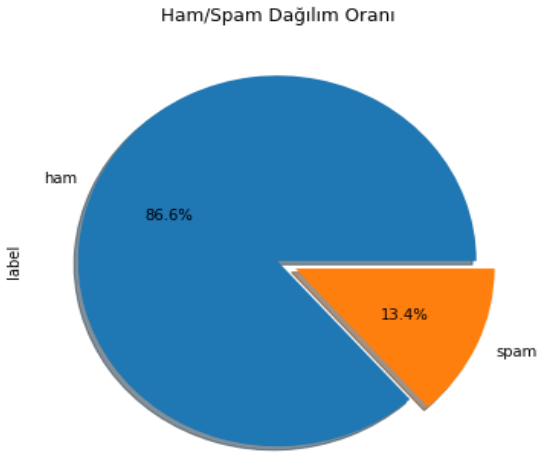
YÖNTEM

1. VERİ SETİNİN TANINMASI

a. Dağılım

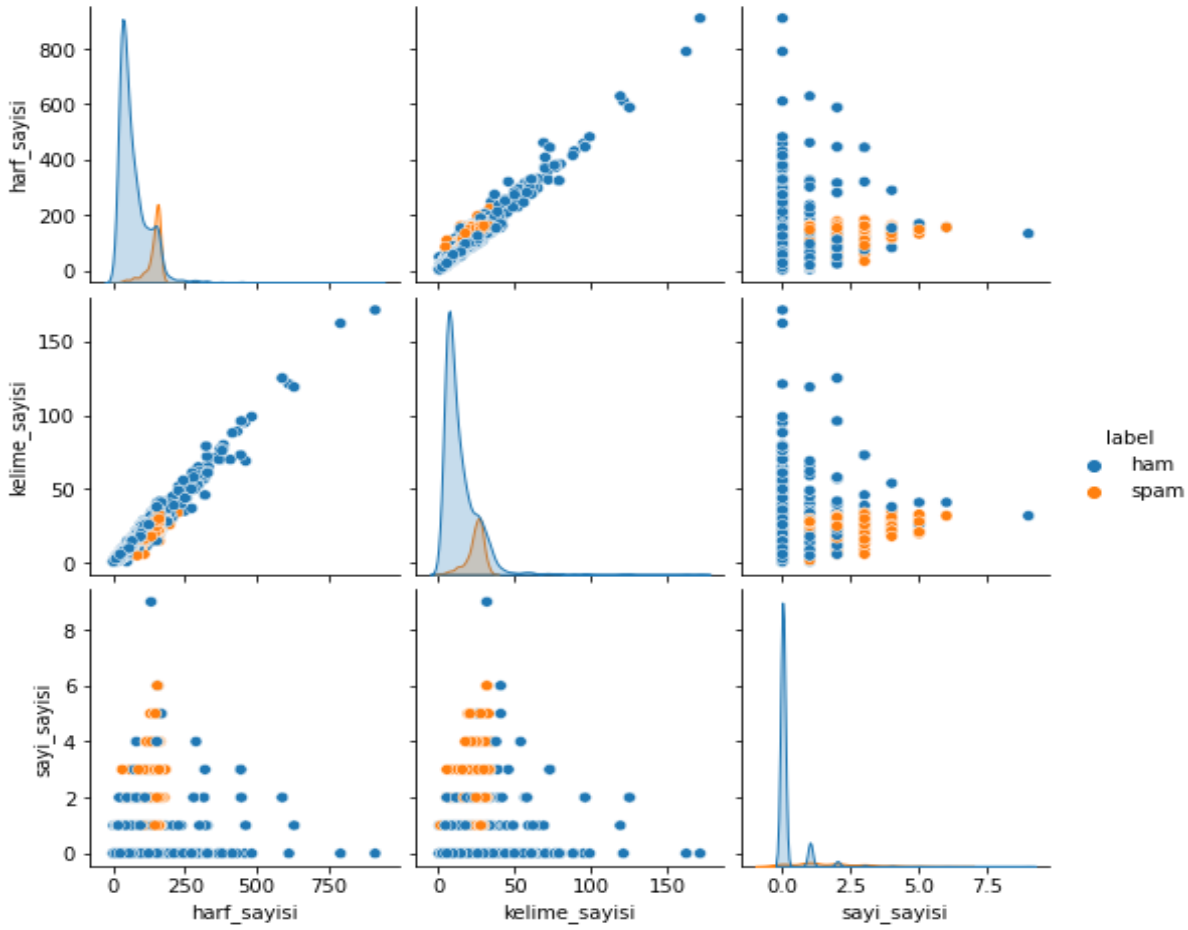
Problemi çözebilmek için öncelikli olarak problemi ve veri setini iyi anlayıp yorumlayabilmek gereklidir. Mesaj ve etiket sütunları altında 5574 değere sahip veri setindeki ham ve spam mesajlarının oran ve dağılım grafikleri aşağıda verilmiştir.





b. Matematiksel İşlemler ve Basit Özellik Çıkarımı

Veri seti içerisindeki ham ve spam mesajlarının bulundurdukları kelime, harf ve rakam sayısından oluşan veri dizisi aşağıda görselleştirilmiştir.



2. METİN/VERİ ÖN İŞLEME

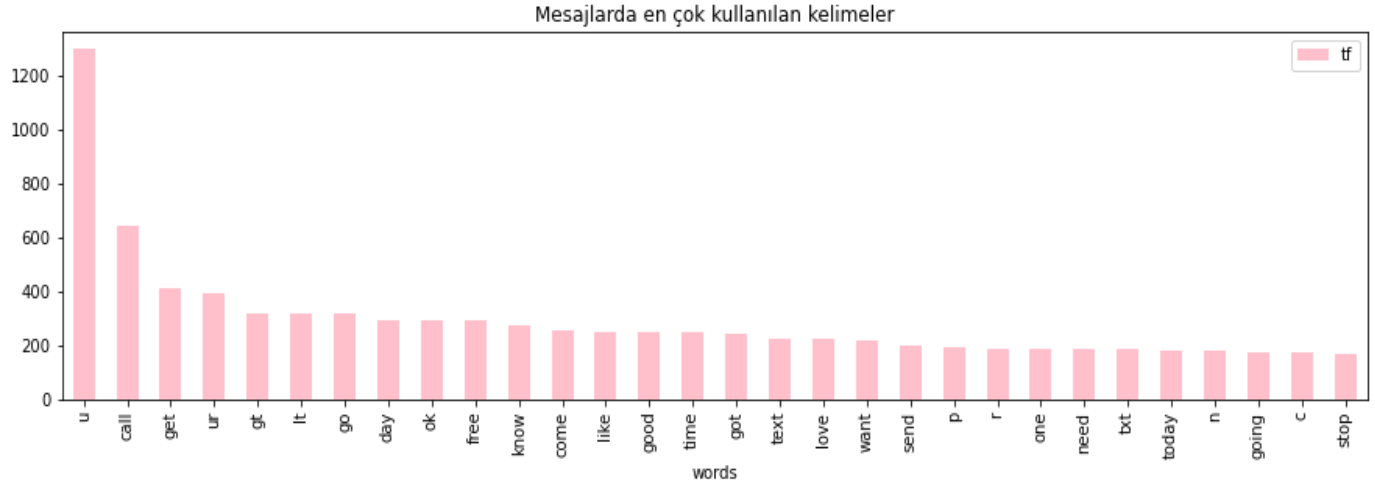
Oluşturulacak makine öğrenmesi modelinin yapısal olmayan verilerden anlam çıkarabilmesi için verilerin işlenebilir hale getirilmesi elzemdir. Verinin işlenebilirlik kıstasını sağlayabilmesi için;

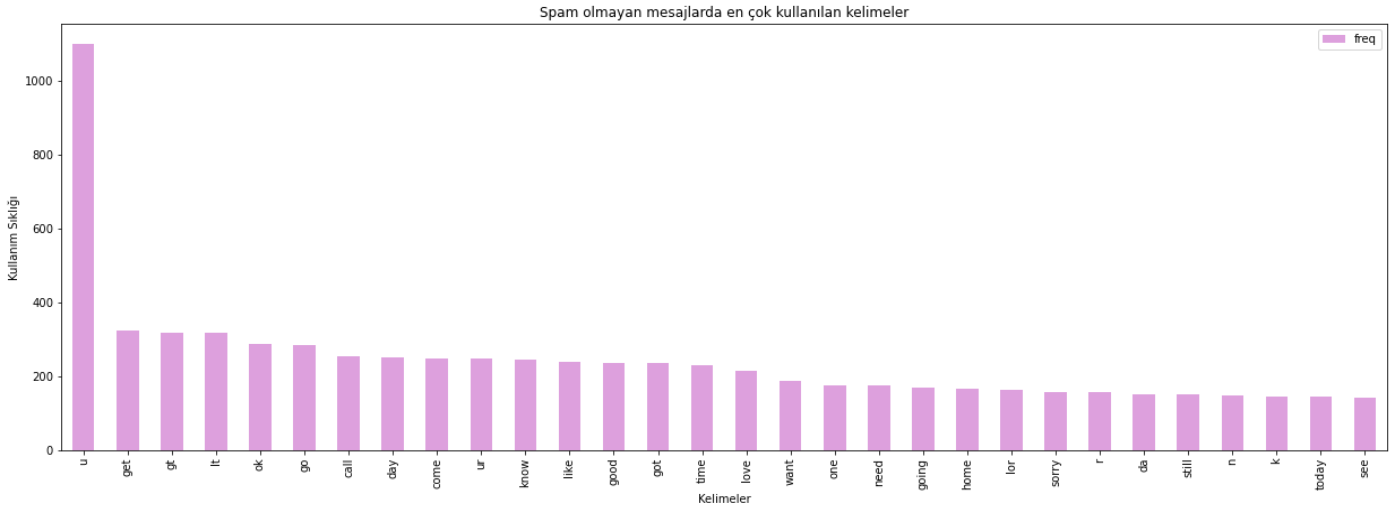
a. **Özel Karakterlerin Ayrımı:** Yalnızca Latin alfabesinde bulunan harflerin kullanımı.
b. **Büyük - Küçük Karakter Dönüşümü:** Aynı harfin büyük - küçük karakterlerinin model tarafından farklı algılanmaması amacıyla tüm karakterlerin küçük harfle ifade edilmesi.
c. **Stop Words:** "I, you, the, at" gibi ifadelerin modelin doğruluğunu bozmaması amacıyla çıkarılması.
ç. **Seyrek Bulunan İfadelerin Silinmesi:** Metinde az kullanılan ifadelerin sınıflandırmayı etkilememesi dolayısıyla silinmesi.
d. **Lemmitization:** Farklı ekler ile çekimlenmiş ifadelerin kök haline indirgenmesi. (goes, going → go)

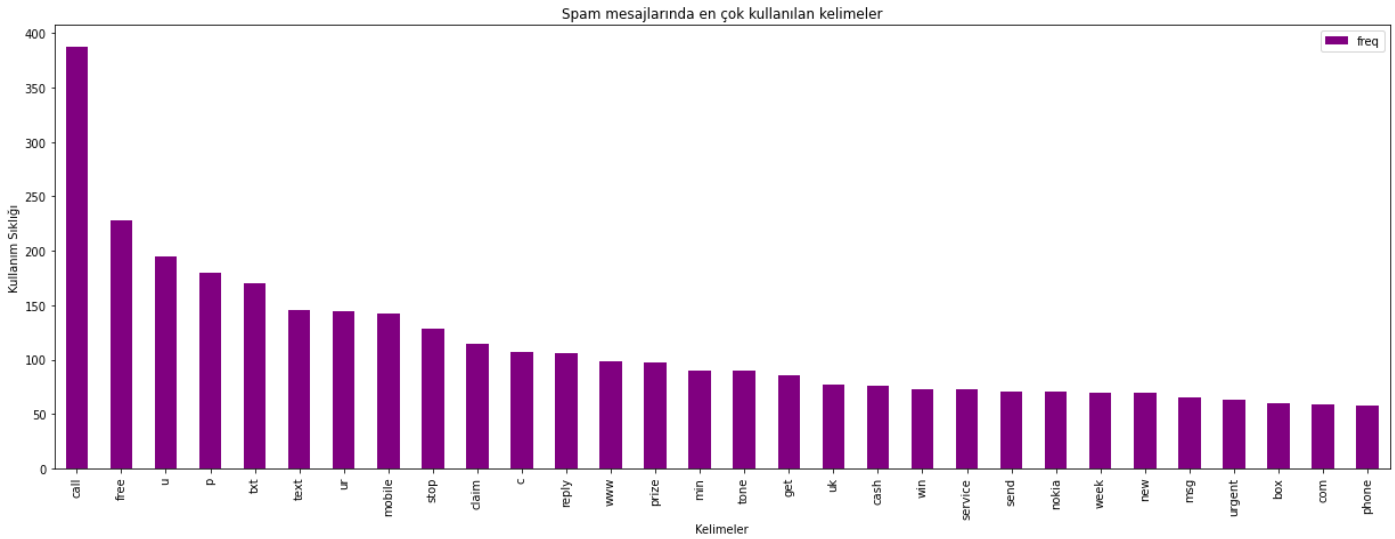
işlemleri uygulanmıştır.

3. TERİM FREKANSI (TERİMLERİN KULLANIM SIKLIĞI)

Bu aşamada, görselleştirme işlemleri öncelikli olarak tüm mesajlarda en çok kullanılan kelimeler, ardından da ham ve spam mesajlarında en çok kullanılan kelimeler olarak gerçekleştirilmiştir.







4. WORD CLOUD İLE SIK KULLANILAN İFADELERİN GÖRSELLEŞTİRİLMESİ

a. Spam Mesajları

b. Ham Mesajları

5. DEĞİŞKEN MÜHENDİSLİĞİ

Model oluşturma aşamasına geçmeden önce, modelin verilen veriyi anlamlandırabilmesi için verinin nümerik hale getirilmesi gerekmektedir. Metin verisinin nümerik olarak temsil edilebilmesi için kullanılan çeşitli yöntemler olmakla birlikte bu çalışmada Count Vectors ve TF-IDF Vectors (words, n-grams, characters) işlemleri gerçekleştirilmiştir.

6. VERİ SETİNİN BÖLÜNMESİ

Modeli eğitmek için kullanılan veriler öncelikle X ve y olmak üzere ikiye ayrılır. X anlam çıkarmak istediğimiz değerleri (bu projede metin verisini), y ise hedef veriyi (bu projede spam ya da ham ifadesini) temsil eder. Sonrasında veriler eğitilecek (train) ve test edilecek olarak belirli bir orana göre bölünür. Model train setleri ile eğitildikten sonra test setleri ile kontrol edilip performans değerleri hesaplanır.

7. MODELİN OLUŞTURULMASI

Proje, tek bir model üzerinden değil çeşitli modellerin farklı vektör teknikleri vasıtasıyla eğitilip karşılaştırılması şeklinde ilerlemiştir. Oluşturulan modeller sırasıyla şunlardır;

1. Logistic Regression
2. Naive Bayes
3. Random Forest
4. XGBoost
5. SVC
6. SGDClassifier

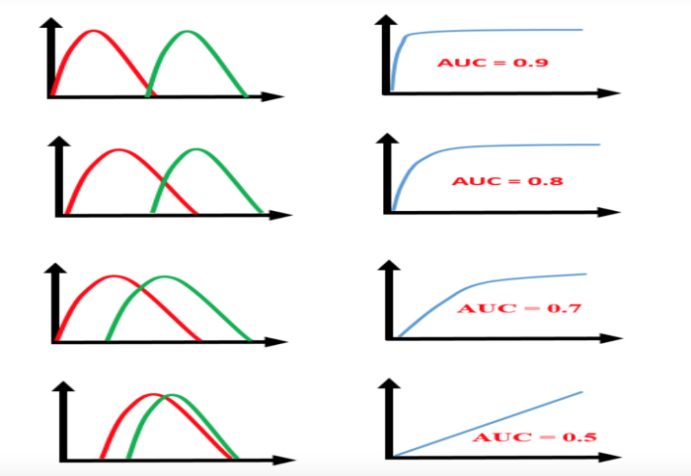
8. PERFORMANS METRİKLERİ

Accuracy değeri performans değerlendirmelerinde sıklıkla kullanılmasına karşın özellikle dengesiz dağılmış bir veri seti özelinde doğru sonuçlar vermeyebilir. Accuracy değeri göz önünde bulundurulmakla birlikte, sınıflandırma problemlerinde performansı tanımlamak amacıyla sıklıkla kullanılan yöntemlerden biri olan Karışıklık Matrisi (Confusion Matrix) değerleri ile hesaplanan farklı metrikler de performans tablosunda yerini almıştır. Karışıklık Matrisinin çalışma mantığı aşağıda bulunan görselde açıklanmaktadır.

Confussion Matrix		Gerçek (Actual) Sonuçlar	
		Pozitif (1)	Negatif (0)
Tahminlenen (Predicted) Sonuçlar	Pozitif (1)	TP [1, 1] True Pozitif	FP [1, 0] False Pozitif
	Negatif (0)	FN [0, 1] False Negatif	TN [0, 0] True Negatif

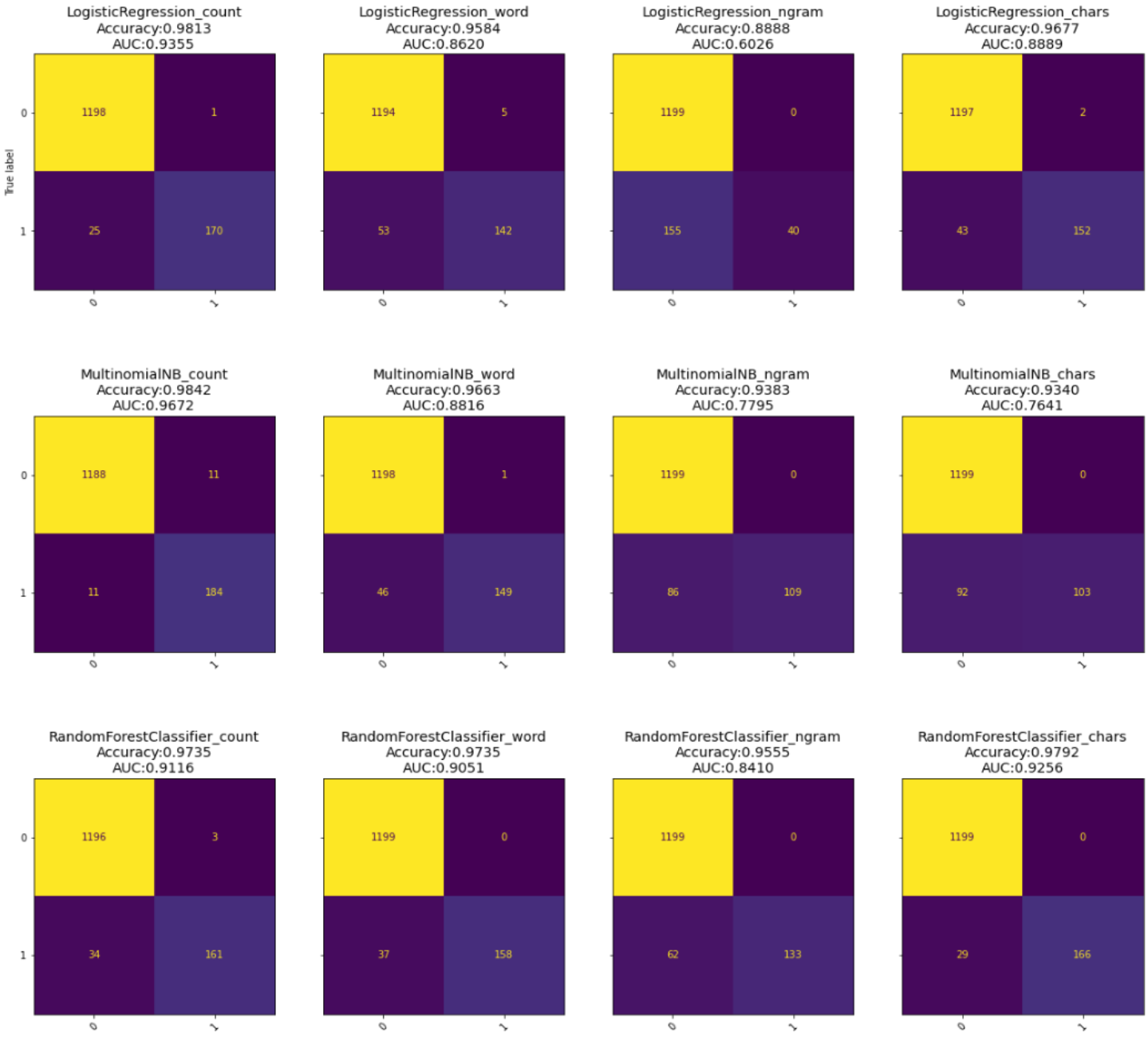
Accuracy = (TP + TN) / (TP + FP + TN + FN)
Precision = TP / (TP + FP)
Recall = TP / (TP + FN)
F1-score = 2 * Precision * Recall / (Precision + Recall)

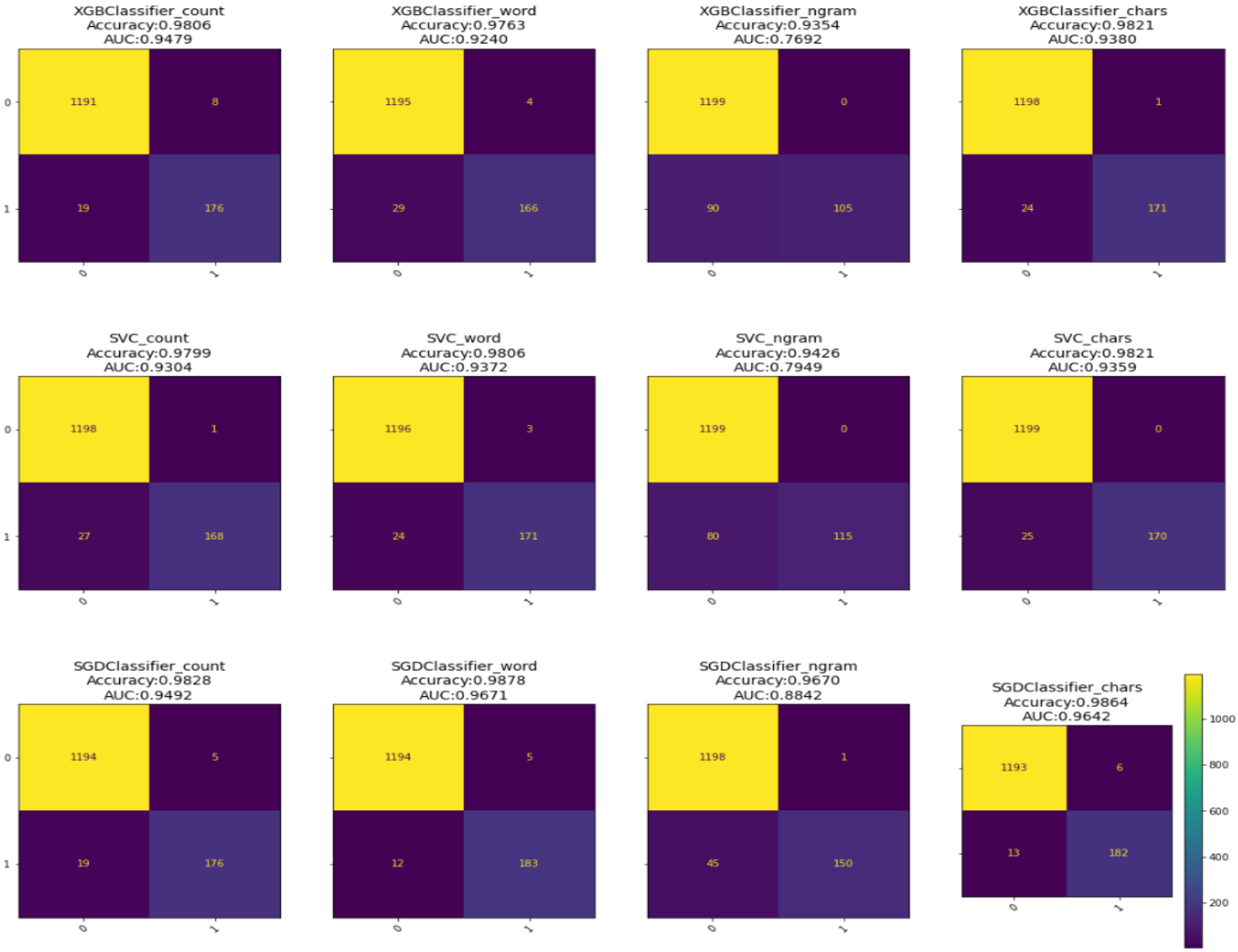
Bir diğer performans metriği olan AUC ise parametrelerin ayrıştırılabilir ölçüsünü veya derecesini temsil eder. AUC metriğinin çalışma mantığı yandaki görselde işlenmiştir. Yeşil ve kırmızı ile temsil edilen eğriler projenin ham ve spam değerleridir. AUC değerinin artması, modelin bu iki sınıfı birbirinden başarılı bir şekilde ayırt edebildiğini gösterir.

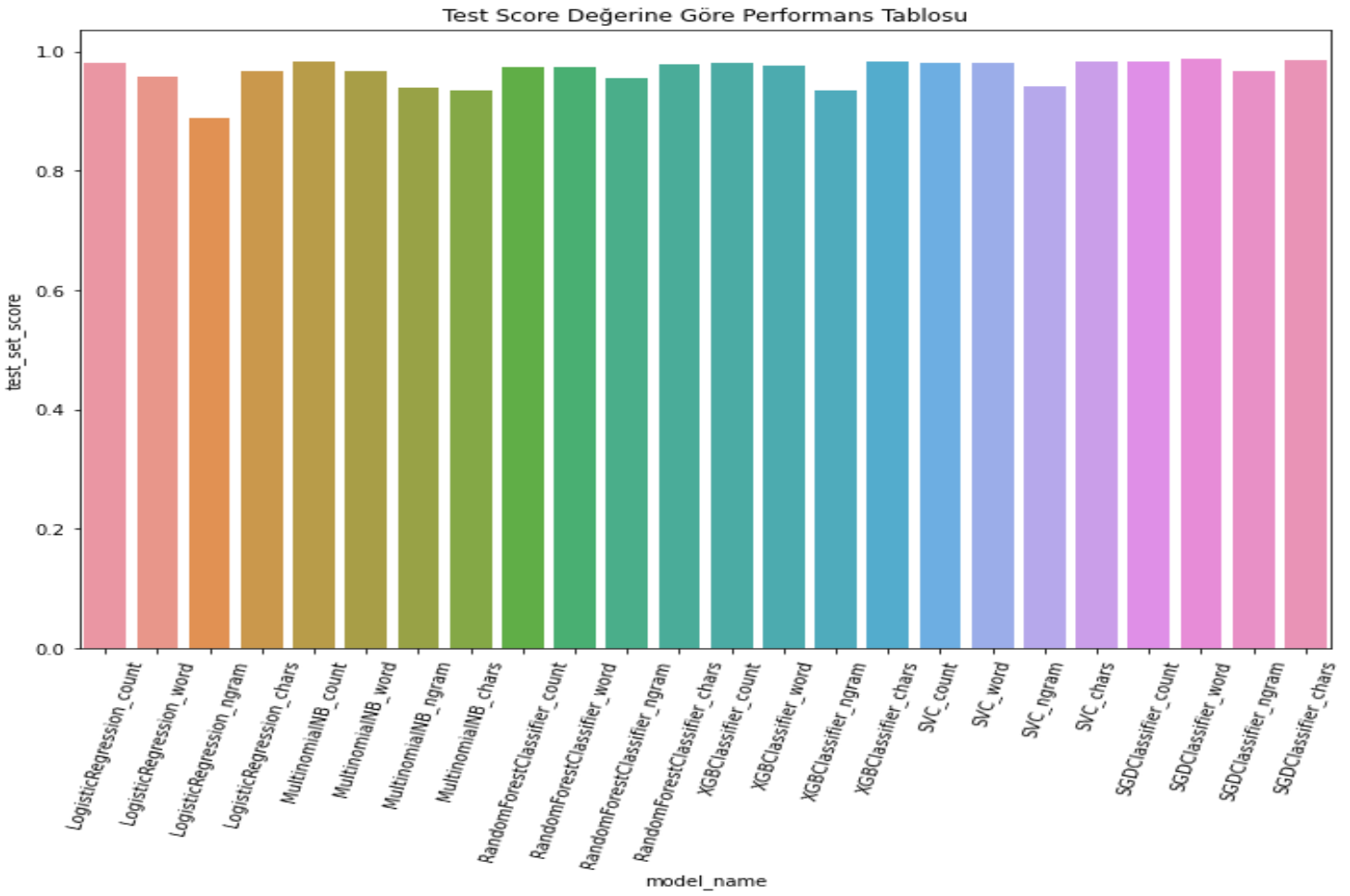


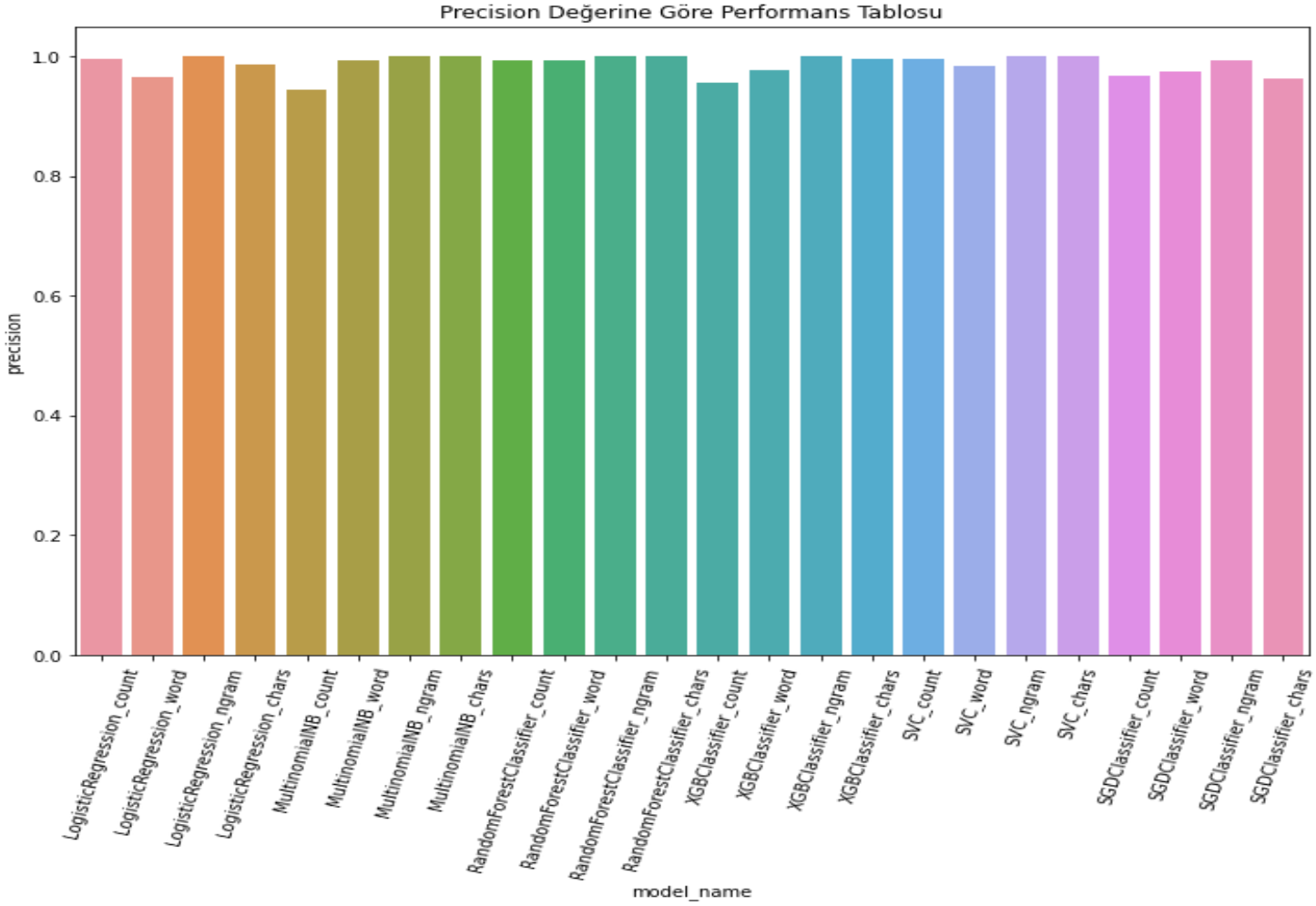
9. PERFORMANS DEĞERLENDİRMESİ

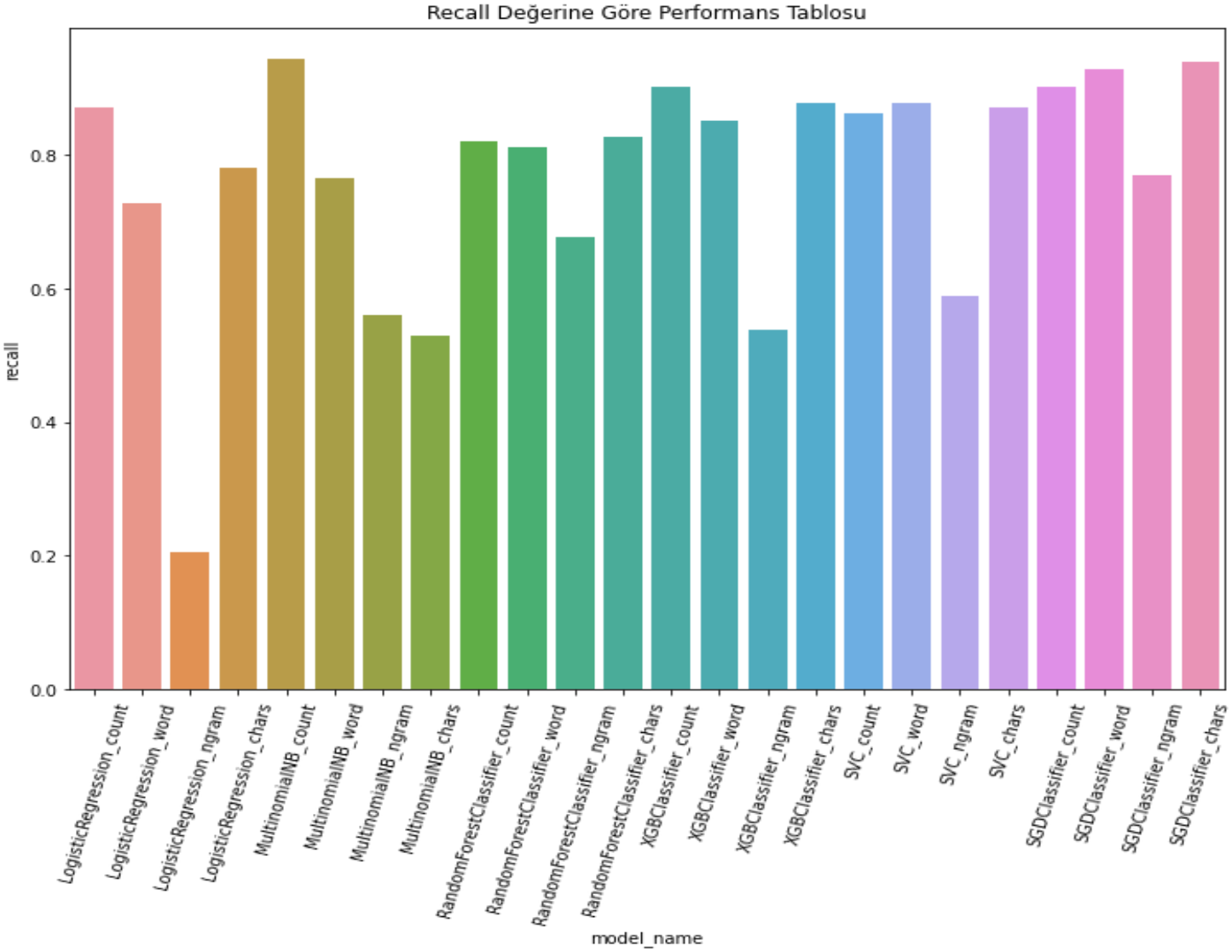
Oluşturulan ve eğitilen modellerin performans değerleri, karmaşıklık matrisleri başta olmak üzere sırasıyla verilmiştir.

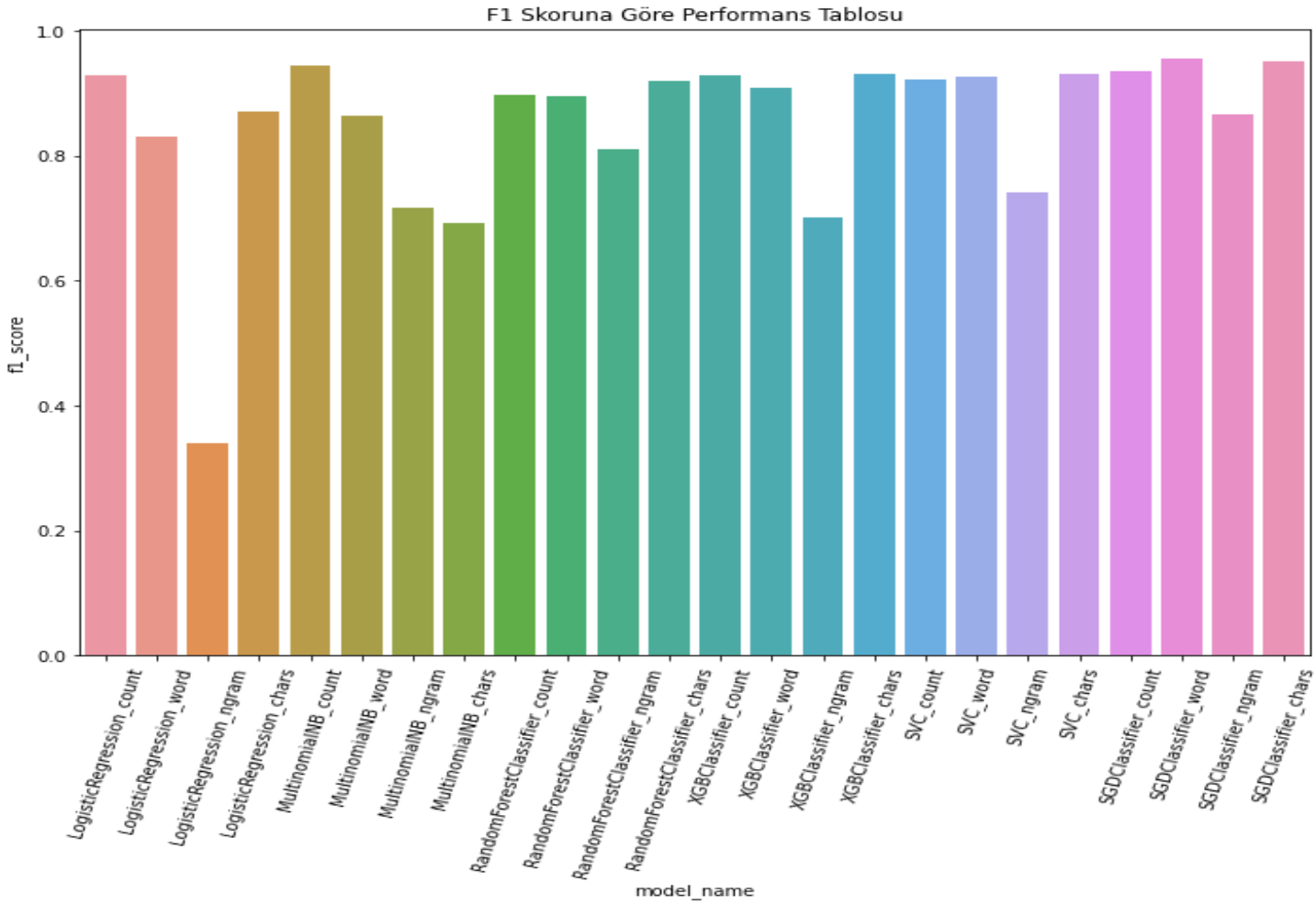


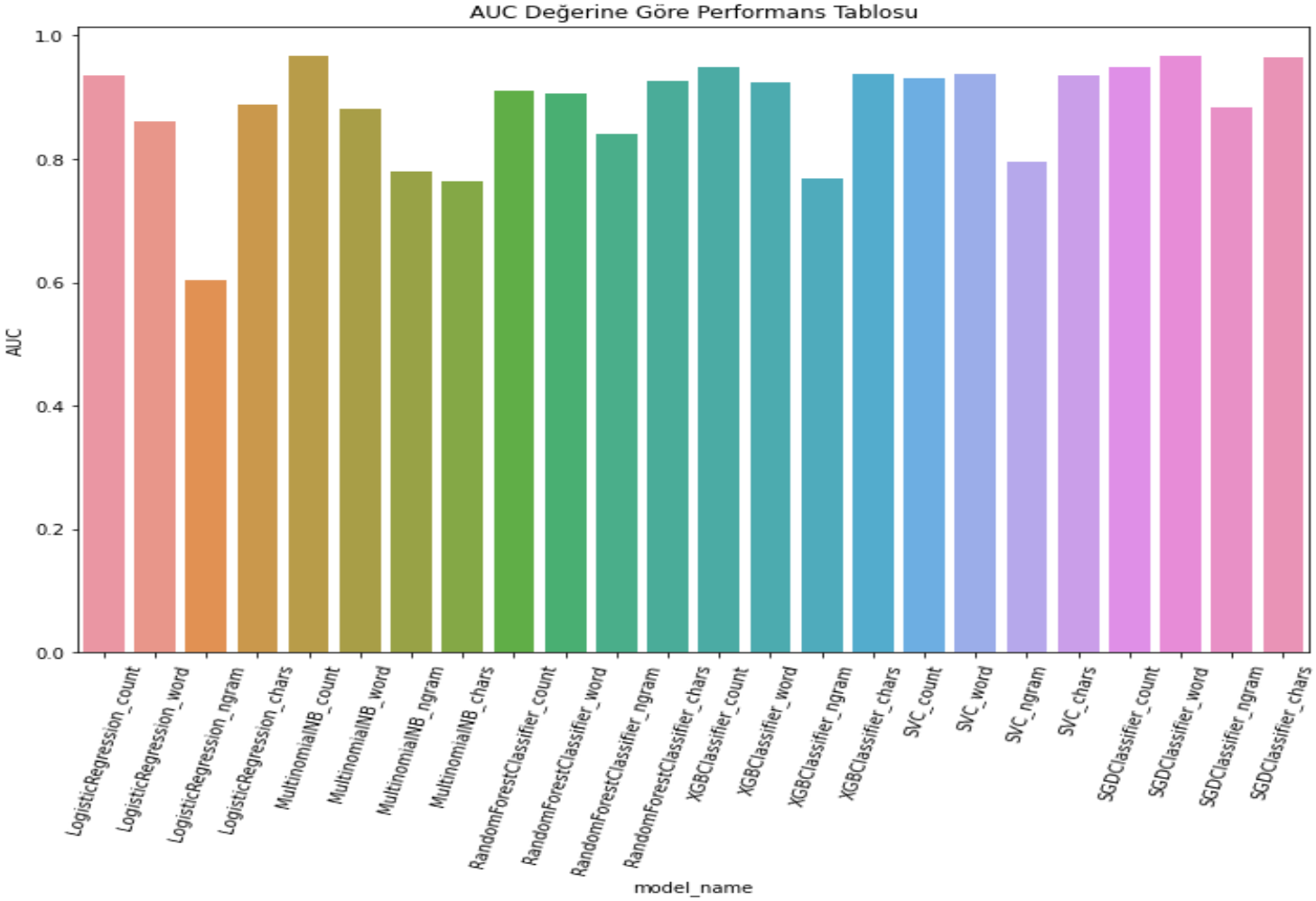

XGBClassifier_count
Accuracy:0.9806
AUC:0.9479
1191
8
19
176
XGBClassifier_word
Accuracy:0.9763
AUC:0.9240
1195
4
29
166
XGBClassifier_ngram
Accuracy:0.9354
AUC:0.7692
1199
0
90
105
XGBClassifier_chars
Accuracy:0.9821
AUC:0.9380
1198
1
24
171
SVC_count
Accuracy:0.9799
AUC:0.9304
1198
1
27
168
SVC_word
Accuracy:0.9806
AUC:0.9372
1196
3
24
171
SVC_ngram
Accuracy:0.9426
AUC:0.7949
1199
0
80
115
SVC_chars
Accuracy:0.9821
AUC:0.9359
1199
0
25
170
SGDClassifier_count
Accuracy:0.9828
AUC:0.9492
1194
5
19
176
SGDClassifier_word
Accuracy:0.9878
AUC:0.9671
1194
5
12
183
SGDClassifier_ngram
Accuracy:0.9670
AUC:0.8842
1198
1
45
150
SGDClassifier_chars
Accuracy:0.9864
AUC:0.9642
1193
6
13
182
0
1
1000
800
600
400
200


Test Score Değerine Göre Performans Tablosu
test_set_score
1.0
0.8
0.6
0.4
0.2
0.0
LogisticRegression_count
LogisticRegression_word
LogisticRegression_ngram
LogisticRegression_chars
MultinomialNB_count
MultinomialNB_word
MultinomialNB_ngram
MultinomialNB_chars
RandomForestClassifier_count
RandomForestClassifier_word
RandomForestClassifier_ngram
RandomForestClassifier_chars
XGBClassifier_count
XGBClassifier_word
XGBClassifier_ngram
XGBClassifier_chars
SVC_count
SVC_word
SVC_ngram
SVC_chars
SGDClassifier_count
SGDClassifier_word
SGDClassifier_ngram
SGDClassifier_chars
model_name


Precision Değerine Göre Performans Tablosu
precision
model_name
LogisticRegression_count
LogisticRegression_word
LogisticRegression_ngram
LogisticRegression_chars
MultinomialNB_count
MultinomialNB_word
MultinomialNB_ngram
MultinomialNB_chars
RandomForestClassifier_count
RandomForestClassifier_word
RandomForestClassifier_ngram
RandomForestClassifier_chars
XGBClassifier_count
XGBClassifier_word
XGBClassifier_ngram
XGBClassifier_chars
SVC_count
SVC_word
SVC_ngram
SVC_chars
SGDClassifier_count
SGDClassifier_word
SGDClassifier_ngram
SGDClassifier_chars


Recall Değerine Göre Performans Tablosu
recall
model_name
LogisticRegression_count
LogisticRegression_word
LogisticRegression_ngram
LogisticRegression_chars
MultinomialNB_count
MultinomialNB_word
MultinomialNB_ngram
MultinomialNB_chars
RandomForestClassifier_count
RandomForestClassifier_word
RandomForestClassifier_ngram
RandomForestClassifier_chars
XGBClassifier_count
XGBClassifier_word
XGBClassifier_ngram
XGBClassifier_chars
SVC_count
SVC_word
SVC_ngram
SVC_chars
SGDClassifier_count
SGDClassifier_word
SGDClassifier_ngram
SGDClassifier_chars


F1 Skoruna Göre Performans Tablosu
f1_score
1.0
0.8
0.6
0.4
0.2
0.0
LogisticRegression_count
LogisticRegression_word
LogisticRegression_ngram
LogisticRegression_chars
MultinomialNB_count
MultinomialNB_word
MultinomialNB_ngram
MultinomialNB_chars
RandomForestClassifier_count
RandomForestClassifier_word
RandomForestClassifier_ngram
RandomForestClassifier_chars
XGBClassifier_count
XGBClassifier_word
XGBClassifier_ngram
XGBClassifier_chars
SVC_count
SVC_word
SVC_ngram
SVC_chars
SGDClassifier_count
SGDClassifier_word
SGDClassifier_ngram
SGDClassifier_chars
model_name


AUC Değerine Göre Performans Tablosu
AUC
1.0
0.8
0.6
0.4
0.2
0.0
LogisticRegression_count
LogisticRegression_word
LogisticRegression_ngram
LogisticRegression_chars
MultinomialNB_count
MultinomialNB_word
MultinomialNB_ngram
MultinomialNB_chars
RandomForestClassifier_count
RandomForestClassifier_word
RandomForestClassifier_ngram
RandomForestClassifier_chars
XGBClassifier_count
XGBClassifier_word
XGBClassifier_ngram
XGBClassifier_chars
SVC_count
SVC_word
SVC_ngram
SVC_chars
SGDClassifier_count
SGDClassifier_word
SGDClassifier_ngram
SGDClassifier_chars
model_name

	model_name	training_set_score	test_set_score	precision	recall	f1_score	AUC
21	SGDClassifier_word	0.999043	0.987805	0.973404	0.938462	0.955614	0.967146
23	SGDClassifier_chars	0.997129	0.986370	0.968085	0.933333	0.950392	0.964165
4	MultinomialNB_count	0.991148	0.984218	0.943590	0.943590	0.943590	0.967208
20	SGDClassifier_count	0.998804	0.982783	0.972376	0.902564	0.936170	0.949197
15	XGBClassifier_chars	1.000000	0.982066	0.994186	0.876923	0.931880	0.938045
19	SVC_chars	0.996411	0.982066	1.000000	0.871795	0.931507	0.935897
0	LogisticRegression_count	0.994976	0.981349	0.994152	0.871795	0.928962	0.935480
12	XGBClassifier_count	0.991148	0.980631	0.956522	0.902564	0.928760	0.947946
17	SVC_word	0.997608	0.980631	0.982759	0.876923	0.926829	0.937210
16	SVC_count	0.995694	0.979914	0.994083	0.861538	0.923077	0.930352
11	RandomForestClassifier_chars	1.000000	0.979197	1.000000	0.851282	0.919668	0.925641
13	XGBClassifier_word	0.992584	0.976327	0.976471	0.851282	0.909589	0.923973
9	RandomForestClassifier_word	1.000000	0.973458	1.000000	0.810256	0.895184	0.905128
8	RandomForestClassifier_count	1.000000	0.973458	0.981707	0.825641	0.896936	0.911569
3	LogisticRegression_chars	0.975359	0.967719	0.987013	0.779487	0.871060	0.888910
22	SGDClassifier_ngram	0.999761	0.967001	0.993377	0.769231	0.867052	0.884198
5	MultinomialNB_word	0.977273	0.966284	0.993333	0.764103	0.863768	0.881634
1	LogisticRegression_word	0.970574	0.958393	0.965986	0.728205	0.830409	0.862017
10	RandomForestClassifier_ngram	0.999761	0.955524	1.000000	0.682051	0.810976	0.841026
18	SVC_ngram	0.998325	0.942611	1.000000	0.589744	0.741935	0.794872
6	MultinomialNB_ngram	0.972727	0.938307	1.000000	0.558974	0.717105	0.779487
14	XGBClassifier_ngram	0.953589	0.935438	1.000000	0.538462	0.700000	0.769231
7	MultinomialNB_chars	0.945455	0.934003	1.000000	0.528205	0.691275	0.764103
2	LogisticRegression_ngram	0.897368	0.888809	1.000000	0.205128	0.340426	0.602564

SONUÇ

Çeşitli işlemler sonrasında işlenebilir hale getirilmiş veri ile eğitilen modellerin pek çok metriğe dayanarak oluşturulmuş performans tablosunda görüleceği üzere, aynı model üzerinden eğitilmelerine karşın farklı vektörizasyon yöntemlerinin kullanımı genel performansı oldukça etkileyebilmektedir. Bununla birlikte, SGDClassifier modelinin genel olarak en iyi performansı gösterdiği söylenebilir.