ICONTECH INTERNATIONAL JOURNAL OF SURVEYS, ENGINEERING, TECHNOLOGY
ISSN 2717-7270
Journal homepage: http://icontechjournal.com/index.php/iij
Volume 5 (2021) Issue 3

AVRUPA TOPLULUĞU EKONOMİK FAALİYETLERİN İSTATİSTİKİ SINIFLANDIRILMASI KULLANILARAK DENGESİZ VERİ SETLERİNDE SINIFLANDIRMA PROBLEMİNE BAKIŞ

AN OVERVIEW OF THE CLASSIFICATION PROBLEM IN UNBALANCED DATASETS USING THE STATISTICAL CONSTRUCTION OF EUROPEAN COMMUNITY ECONOMIC ACTIVITIES

Yasin BEKTAŞ

Mersin Üniversitesi, Erdemli Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Erdemli, MERSİN

ORCID ID: 0000-0002-2761-5780

Jale BEKTAŞ

Mersin Üniversitesi, Erdemli Uygulamalı Teknoloji ve İşletmecilik Yüksekokulu, Bilgisayar Teknolojisi ve Bilişim Sistemleri Bölümü, Erdemli, MERSİN

ORCID ID: 0000-0002-8793-1486

ÖZET

Dengesiz ve çok sınıflı veri setlerinde klasik sınıflandırıcıların kullanılması her zaman bir sorun oluşturmuştur. Bu çalışmada Avrupa Topluluğunda Ekonomik Faaliyetlerin İstatistiki Sınıflaması (NACE) kodlarının tanımları üzerinde çok bilinen sınıflandırıcılar ile bir metin madenciliği uygulaması yapılmıştır. Çalışmada öncelikle orjinal verinin dengesiz yapısı üzerinde uygulama yapılmış, daha sonra sınıf bazında ağırlıklandırma yöntemiyle dengeli hale getirilerek sonuç verisi üzerinde tekrar test edilerek performans ölçümü gerçekleştirilmiştir. Testlerde Karar Ağaçları, Naiv Bayes, Destek Vektör Makineleri, Çapsal Tabanlı Fonksiyonlar ve Rastgele Orman algoritmaları gibi yaygın kullanılan sınıflandırıcılar kullanılmıştır. Çalışma bize Karar Ağaçlarının veri dengelenmesi neticesinde F-skor değerinin %17.43' den %92' ye çıkarak en iyi performansı verdiğini göstermiştir.

Anahtar Kelimeler: Metin madenciliği, Dengesiz veri seti, Sınıflandırıcılar, Nace

ABSTRACT

The use of classical classifiers in unbalanced and multi-class data sets has always been a problem. In this study, a text mining work has been applied with well-known classifiers on the definitions of Statistical Construction of Economic Activities (NACE) codes in the European Community. In the study, first of all, the application was made on the unbalanced structure of the original data, then the performance measurement was performed by retesting the result data by making it balanced by weighting on a class basis. Common classifiers such as Decision Trees, Naiv Bayes, Support Vector Machines, Diametric Based Functions and Random Forest algorithms were used in the tests. The study showed us that as a result of data balancing of Decision Trees, the F-score value increased from 17.43% to 92%, giving the best performance.

Keywords: Text mining, Unbalanced dataset, Classifiers, Nace

GİRİŞ

Avrupa Topluluğunda Ekonomik Faaliyetlerin İstatistiki Sınıflaması (NACE); ekonomik faaliyetlerin Avrupa'da istatistiksel değerlendirmelerle ortaya konulması ve bunun yayılmasını sağlamak amacıyla hazırlanmış ve detaylandırılmış bir sistemdir(Nace, 2008). Bir kodlama sistemine sahiptir ve faaliyet kodlarına göre sınıflandırılan bu sistem çerçevesinde işletmelere detay iştigaline göre altı haneli bir kod verilmektedir. Açılımı "Nomenclature des Activités Économiques dans la Communauté Européenne" şeklindedir ve ismini bunun baş harflerinden alır. NACE kodu uygulaması, "Tüm Ekonomik Faaliyetlerin Uluslararası Standart Sanayi Sınıflaması (ISIC)" ile bağlantılıdır ve ekonomik faaliyetlere ilişkin istatistiki verileri dünya standartlarında karşılaştırma açısından da oldukça önemli bir araçtır(Uyumsoft, 2020). Türkiye Cumhuriyeti Ticaret Bakanlığı tarafından belirlenip, revizyonları TOBB tarafından sunulmaktadır. Bu araştırmada, NACE kod sisteminin detay kayıtlarının açıklamaları ve bu detay bilgilerinin ait olduğu sınıf bilgileri veri seti olarak kullanılmıştır. Çalışmada da faaliyet tanımlarının hangi sınıfa ait olduğunun tespiti farklı sınırlayıcılar yapılmıştır.

Metin madenciliği uygulamaları bir süreçtir ve bu sürecin başarılı bir şekilde yürütülmesiyle olumlu sonuçlara ulaşılabilmektedir (Berry, 2004). Metin verisi içeren veri setlerinde metin madenciliğinin her aşamasında

Received 17 August 2021; Received in revised form 25 August 2021; Accepted 21 September 2021;
Available online 25 September 2021;
doi: 10.46291/ICONTECHvol5iss3pp31-37

kullanılan dilbilgisel, istatistiksel teknik ve algoritmaların incelenmesi ve karşılaştırılması gerekmektedir(Agrawal ve Batra, 2013). Önişleme sürecinde parçalara ayırmak (tokenization) işlemi uygulanarak cümledeki kelimelerin bulunması ve kelimelerin diğer kelimelerle ilişkisine bakılmaktadır. Diğer bir işlem ise kelimelerin yapısını belirleme işlemi (Pos tags) yapılarak duruma göre uygun kelime etiketleri belirlenmesidir. Veri setindeki etkisiz kelimelerin çıkarımı (Stop words) yapılarak "ve, veya, ile, ise" gibi veri seti tanımlamada yardımcı olmayan kelimeler ve noktalama işaretleri çıkarılır(Juson ve Alfawareh, 2012).

Önişlemden geçirilen bu türdeki veri setleri üzerinden pek çok yöntemle sınıflandırma yapılmıştır. Çalışmalardan birinde, Naive Bayes Sınıflayıcı (NBC), K-En Yakın Komşu (KNN), and Karar Ağacı (DT) kullanılarak sosyal medyada kullanılan Endonezce sövgü kelimelerinin tespiti yapılmıştır(Zulfikar ve ark., 2017). Bir başka çalışmada Karar Destek Makineleri (SVM) sınıflandırma modeli, haber tabanlı haber referanslarını kullanarak ekonomi politikası belirsizliğini ölçmek için tercih edilmiştir (Tobback ve ark., 2018). Film yorumlarının içeriğine göre Naive Bayes, Merkez Tabanlı Sınıflayıcı, Çok Katmanlı Yapay Sinir Ağları (MLP) ve Destek Vektör Makineleri (SVM) gibi makine öğrenmesi yöntemleri kullanılarak duygu düşünce analizi bir başka çalışmada yer almıştır (Kaynar ve ark., 2016).

Çalışmada kullanılan NACE veri seti üzerinde Almanya kaynaklı ekonomik hareketlerde istatistiksel yöntemlerle sınıflandırma yapılmıştır (Schnabl ve Zenker, 2013). Bir başka çalışmada NACE altındaki alanlar için zaman serilerini yeniden oluşturmak için kullanılabilecek farklı örnekleme ve tahmin stratejileri açıklanmıştır (Van den Brakel, 2010).

Bu çalışmada ise ön işleme yöntemleri ile hazırlanan Nace detay bilgileri Destek Vektör Makineleri(SVM), Naive Bayes Sınıflandırıcı (NBC), Karar Ağaçları(DTC), Rastgele Orman Algoritması(RFC) ve Çapsal Tabanlı Fonksiyonlar (RBF) sınıflandırıcıları ile sınıflandırılma başarıları incelenmiştir. Bu işlemler dengesiz özellikteki orijinal veri seti ile ve sınıf ağırlıkları dengelenmesi yöntemi ile veri seti dengelendikten sonra ayrı ayrı test edilmiş ve başarımları karşılaştırılmıştır.

1. METOD

1.1. SINIFLANDIRICILAR

1.1.1. DESTEK VEKTÖR MAKİNELERİ (SUPPORT VEKTOR MACHİNE - SVM)

Destek Vektör Makineleri sınıflandırma işlemini yaparken veri içindeki farklı sınıflara ait en yakın örneklerin ayırıcı yüzeylere dik olarak ölçülen mesafelerini maksimize etmeyi amaçlar. SVM' nin diğer birçok sınıflandırıcılardan olan üstünlüğü sınıf ayırıcı yüzeyin her iki sınıfa da olan uzaklığı aynı maksimum değerde olmasıdır. Sınıflar arası boşluğun doğrusal olmadığı durumlarda uygulanabilecek bir çekirdek fonksiyon yardımıyla bir üst uzaya haritalanan veriye doğrusal SVM çözümü uygulanabilir. Çalışmada çekirdek fonksiyonu olarak Çapsal Tabanlı Fonksiyonlar olarak isimlendirilen RBF fonksiyonu kullanılmıştır. Formül detayı ise ;

$$K(x_i, x_j) = exp\left(-\frac{1}{2}\|x_i - x_j\|^2\right)$$

1.1.2. NAİV BAYES SINIFLANDIRICILARI(NAİVE BAYES CLASSİFİER - NBC)

Naive bayes yöntemi bayes teoremine dayanan bir sınıflandırma yöntemidir. Bu sınıflandırıcının temel noktası veri setindeki tüm nitelikleri birbirinden bağımsız olarak hesaplamasıdır. Bayes teoreminin bir olayın gerçekleşme ihtimalinin başka bir olayın gerçekleşmesine olan bağlantısının formülü;

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

olarak ifade edilir. Bayes teoreminden yola çıkarak her örneğin ait olduğu sınıf hesaplamasında kullanılan formül;

$$P(x|S_i) = P(S_i)\prod_{k=1}^{L} P(x_k|Si)$$

Burada L nitelik S ise sınıf değerlerini temsil etmektedir. Hesaplama sonucunda x örneğinin ihtimali en yüksek sınıf oranı hesaplanarak bulunmaktadır.

1.1.3. KARAR AĞAÇLARI SINIFLANDIRICISI (DECİSİON TREE CLASSİFİER - DTC)

Karar ağaçlarında amaç, veriyi bilgi kazancının en yüksek olduğu niteliklerden bölümlere ayırmaktır. Çalışmada ID3 ten geliştirilmiş olan C4.5 kazanım oranı(Gain Ratio) kullanılmıştır.

Received 17 August 2021; Received in revised form 25 August 2021; Accepted 21 September 2021;
Available online 25 September 2021;
doi: 10.46291/ICONTECHvol5iss3pp31-37

1.1.4. RASTGELE ORMAN ALGORTİMASI(RANDOM FOREST CLASSİFİER - RFC)

Rastgele Orman Algoritmasında amaç, karar ağaçlarındakine benzer biçimde rastgele bir şekilde veri setinden alt kümeler alınarak birden fazla ağaç üretmek ve sonucunda sınıflandırma başarısını yükseltmektir. Kullanılan ağaç sayısı arttıkça hata oranının düştüğü bu algoritma aynı zamanda çok fazla niteliğe ve sınıf sayısına sahip dengesiz veri setlerinde de oldukça yüksek başarı değerleri yakalayabilmektedir.

1.1.5. ÇAPSAL TABANLI FONKSİYONLAR (RADİAL BASİS FUNCTİON - RBF)

İleri beslemeli bir sinir ağı çeşidi olan RBF ağlarında eğiticili öğrenme yapılmaktadır. RBF ağlarında eğitim esnasında sistem öncelikle gizli katmandaki nöron sayısını ve çıkış ağırlıklarını hesaplar. Yöntemlerin neredeyse tamamında nöron sayısı deneyerek elde edilmektedir.

1.2. VERİ SETİ

Çizelge 1: Nace Kodlama Sistemi Sınıf Detay Bilgileri

Sınıf Adı	Sınıf Açıklama	Örnek Sayısı	Sınıf Ağırlık
C	İmalat	913	0.42
G	Toptan Ve Perakende Ticaret; Motorlu Kara Taşıtlarının Ve Motosikletlerin Onarımı	440	0.2
H	Ulaştırma Ve Depolama	112	0.05
A	Tarım, Ormancılık Ve Balıkçılık	85	0.04
S	Diğer Hizmet Faaliyetleri	78	0.04
F	İnşaat	69	0.03
N	İdari Ve Destek Hizmet Faaliyetleri	68	0.03
M	Mesleki, Bilimsel Ve Teknik Faaliyetler	67	0.03
R	Kültür, Sanat, Eğlence, Dinlence Ve Spor	48	0.02
B	Madencilik Ve Taş Ocakçılığı	45	0.02
J	Bilgi Ve İletişim	40	0.02
Q	İnsan Sağlığı Ve Sosyal Hizmet Faaliyetleri	38	0.02
K	Finans Ve Sigorta Faaliyetleri	36	0.02
I	Konaklama Ve Yiyecek Hizmeti Faaliyetleri	35	0.02
P	Eğitim	35	0.02
O	Kamu Yönetimi Ve Savunma; Zorunlu Sosyal Güvenlik	32	0.01
D	Elektrik, Gaz, Buhar Ve İklimlendirme Üretimi Ve Dağıtımı	15	0.01
E	Su Temini; Kanalizasyon, Atık Yönetimi Ve İyileştirme Faaliyetleri	15	0.01
L	Gayrimenkul Faaliyetleri	0	0
T	Hane halklarının İşverenler Olarak Faaliyetleri; Hane halkları Tarafından Kendi Kullanımlarına Yönelik Olarak Ayrım Yapılmamış Mal Ve Hizmet Üretim Faaliyetleri	0	0
U	Uluslararası Örgütler Ve Temsilciliklerinin Faaliyetleri	0	0
V	Kendi Adına Menkul Sermaye İradı Faaliyetleri (Temettü, Banka Faizi, İştirak Kazançları Vb.)	0	0
	Toplam	**2171**	**1**

Veri seti girişte bahsedilen NACE kodlama sistemi temel alınarak oluşturulmuştur. Detaylandırılmış halde Çizelge 1 de görülen veri seti toplamda 22 sınıfa ait iken örnek sayısı 10 ve altında olan sınıflar çalışma dışında tutulmuşlardır. Bunun neticesinde 18 adet sınıf ve 2171 örnek ile testler gerçekleştirilmiştir.

Received 17 August 2021; Received in revised form 25 August 2021; Accepted 21 September 2021;
Available online 25 September 2021;
doi: 10.46291/ICONTECHvol5iss3pp31-37

Tüm sınıflardaki Nace detay açıklamaları birkaç ön işlemden geçirilmiştir. Bunlardan ilki noktalama işaretleri ve özel semboller temizlendikten sonra tokenization olarak bilinen alt birimlere ayrıştırma işlemi. Buna bir nevi ifadeleri kelimelerine ayırma işlemi de denilebilir.

Alt birimlere ayırma işleminin devamında metne tek başına anlamsal bir değer katmayan stop-word olarak bilinen kelime grupları bu veriden temizlenmiştir. Bu işlem daha sonra oluşacak nitelik sayısının azalmasını ve işlem yükününde belirgin ölçüde düşmesini sağlamaktadır.

Bilgi getirimi (Information Retrieval -IR) konularında sıkça kullanılan Terim Frekansı-Ters Doküman Frekansı (Term Frequency – Inverse Document Frequency – TF-IDF) değerleri kullanılarak metin içeriğe sahip her bir Nace detay açıklaması sayısal bir veriye dönüştürülmüştür. TF-IDF hesabında TF(Terim Frekansı) değeri ve IDF(Ters Doküman Frekansı) değerleri ayrı ayrı hesaplanır ve çarpılır. Terim frekansı değeri;

$$tf_{i,d} = fr_{i,d}/dfr$$

Formülü ile hesaplanır. Yani *i* teriminin *d* dokümanındaki frekansının *d* dokümanındaki en yüksek tekrar sayısına sahip teriminin frekansına oranıdır. Daha sonra IDF hesabı ise;

$$idf_i = \log(n/df_i)$$

Formülü ile hesaplanır. Burada ise herhangi bir i terimi için toplam doküman sayısının terimi içeren doküman sayısına oranının logaritması ile hesaplanır. Burada dikkat edilmesi gereken konulardan bir tanesi de paydanın sıfır olma olasılığıdır. Daha sonra bu TF ve IDF değerleri her bir dokümandaki her bir terim için çarpılarak metin içerikli veri seti sayısallaştırılmıştır.

Tüm dokümanlar alt birimlerine ayrıştırıldıktan ve stop-word temizliği yapıldıktan sonra 6634 niteliğe sahip bir veri seti elde edilmiştir. Daha öncede bahsedildiği üzere bu veri setimiz sınıfsal bir dengesizliğe sahiptir. Çizelge 1'de de gösterildiği gibi toplam 18 sınıftan "C" sınıfında 913 örnek var iken "D" ve "E" sınıflarında 15' er örnek bulunmaktadır. Testlerde bu şekliyle kullanıldığı gibi sınıf ağırlıklarını dengeleyici bir yeni nitelik eklediğimiz ikinci bir veri seti ile daha test yapılmıştır. Sınıf ağırlığını dengelemek adına her sınıf örneğine aşağıdaki formül kullanılarak yeni bir nitelik eklenmiştir.

$$x_i = K/(C * C_i)$$

Formülde i sınıf indeksini belirtmekte ve x_i ise belirtilen ilgili sınıf için üretilen yeni nitelik değerini belirtmektedir. K veri setindeki toplam kayıt sayısı iken C sınıf sayısı ve C_i ise aynı sınıftaki örnek sayısını göstermektedir. Buna göre kullandığımız veri setindeki dengesizliği ortadan kaldırmak adına oluşturulan yeni ağırlık nitelik değerlerimiz Çizelge 2'de gösterilmiştir.

Çizelge 2:Sınıflar için oluşturulan ağırlık değerleri

Sınıf Adı	Hesaplanan Ağırlık Niteliği
A	1.41895424836601
B	2.68024691358024
C	0.132104174272848
D	8.04074074074074
E	8.04074074074074
F	1.74798711755233
G	0.274116161616161
H	1.07688492063492
I	3.44603174603174
J	3.01527777777777
K	3.3503086419753
M	1.80016583747927
N	1.77369281045751
O	3.76909722222222
P	3.44603174603174
Q	3.17397660818713
R	2.51273148148148
S	1.54629629629629

Received 17 August 2021; Received in revised form 25 August 2021; Accepted 21 September 2021;
Available online 25 September 2021;
doi: 10.46291/ICONTECHvol5iss3pp31-37

2.1. BAŞARI DEĞERLENDİRMESİ

Çalışmada kullanılan veri setinin orijinal hali aşırı dengesiz olmasından dolayı doğruluk(accuracy) değeri bize çok sağlıklı bilgi vermemektedir. Toplam doğru sınıflanan örneklerin tüm örneklere oranı şeklinde tanımlanan doğruluk işlemi eğitim esnasında örnek sayısı fazla olan sınıfa yapılan yoğunluklu tahminde başarıyı yanıltıcı bir şekilde yüksek çıkarmaktadır. Bu sebeple testlerimizde verinin dengeli ve dengesiz durumlarını karşılaştırmak adına doğruluk hesaplamasının yanında duyarlılık(recall) ve kesinlik(precision) metrikleri kullanılarak hesaplanan f-skoru metriği de kullanılmıştır. Bu hesaplamalar için Çizelge 3'de gösterilen karışıklık matrisi kullanılır.

Çizelge 3: Karışıklık Matrisi

Gerçek			
Pozitif	**Negatif**		
Gerçek Pozitif	Yanlış Pozitif	**Pozitif**	**Tahmin Edilen**
Yanlış Negatif	Doğru Negatif	**Negatif**	

Çizelge 3 teki ifadeler kullanılarak duyarlılık değeri;

$$duyarlılık(recall) = \frac{Gerçek\ Pozitifler}{Gerçek\ Pozitifler + Yanlış\ Negatifler}$$

Formülü ile hesaplanır. Aynı şekilde kesinlik değeri ise;

$$kesinlik(precision) = \frac{Gerçek\ Pozitifler}{Gerçek\ Pozitifler + Yanlış\ Pozitifler}$$

Formülü ile hesaplanır. Bu hesaplamaların ardından ise f-skor değeri;

$$f - skor = 2 * \frac{duyarlılık * kesinlik}{duyarlılık + kesinlik}$$

Formülü kullanılarak hesaplanır. Son olarak belirtilmesi geren bir diğer husus ise; çalışmadan toplam 18 sınıf kullanılmasından dolayı yukarıda bahsedilen duyarlılık ve kesinlik değerleri her sınıf için yapılmış ve ortalama değerleri kullanılarak f-skor değeri hesaplanmıştır.

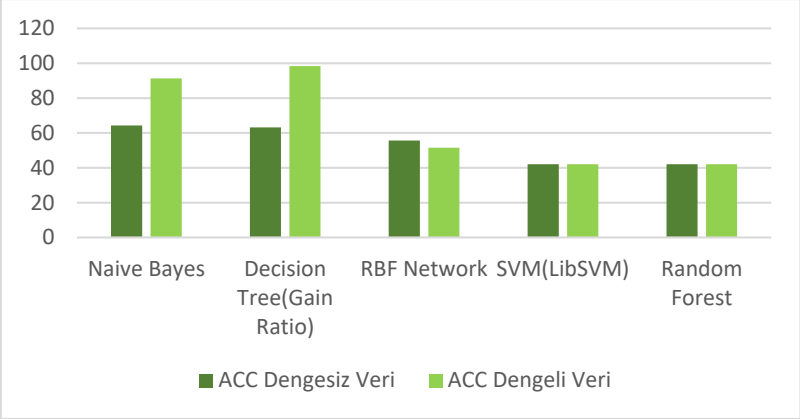
3. SONUÇLAR

Nace kodlama sistemi detay değerlerinin sınıf tahmininde kullanılan yöntemlerin sonuçları iki ayrı tablo şeklinde sunulmaktadır. Öncelikle Çizelge 4'de veri setinin orijinal yani dengesiz hali ile ağırlıklandırma niteliği eklenerek elde edilen dengeli halinin doğruluk yani accuracy değerleri karşılaştırılmıştır. Şekil 1' de ise bu aynı sonuçlar grafiksel olarak sunulmuştur.

Çizelge 4: Her iki verisetinin doğruluk (accuracy) değerleri ile karşılaştırılması.

Yöntem Adı	**ACC Dengesiz Veri**	**ACC Dengeli Veri**
Naive Bayes	**64.29**	91.36
Decision Tree(Gain Ratio)	63.25	**98.39**
RBF Network	55.65	51.5
SVM(LibSVM)	42.05	42.05
Random Forest	42.05	42.05

Received 17 August 2021; Received in revised form 25 August 2021; Accepted 21 September 2021;
Available online 25 September 2021;
doi: 10.46291/ICONTECHvol5iss3pp31-37

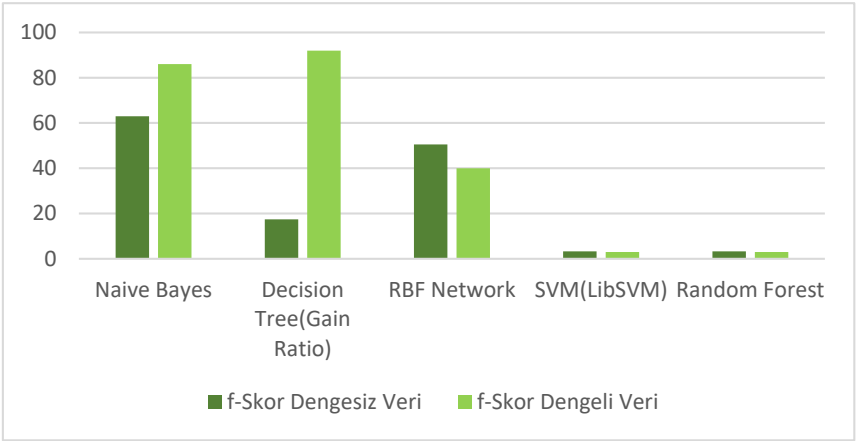


Şekil 1: Her iki verisetinin doğruluk (accuracy) değerleri

Çizelge 5' de ise yine her iki veri seti için f-skor karşılaştırmaları yapılmıştır. Özellikle dengesiz veri setlerinde sınıflandırma başarısı için daha sağlıklı bilgi veren f-skor değerlerinde dengelenen veri seti için bazı yöntemlerde başarı değerinin belirgin arttığı gözlenmiştir. Şekil 2'de bu değerler grafik üzerinde daha net gösterilmiştir.

Çizelge 5: Her iki verisetinin f-skor değerleri ile karşılaştırılması.

Yöntem Adı	**f-Skor Dengesiz Veri**	**f-Skor Dengeli Veri**
Naive Bayes	**62.94**	86
Decision Tree(Gain Ratio)	17.43	**92**
RBF Network	50.52	40
SVM(LibSVM)	3.29	3
Random Forest	3.29	3



Şekil 2: Her iki verisetinin f-skor değerleri

4. TARTIŞMA

Yapılan çalışma neticesinde elde edilen sonuçlar incelendiğinde sınıf ağırlıklarının dengelenmesi Naiv Bayes (NBC) ve Karar Ağaçları (DTC) algoritmalarının başarıyı olumlu yönde değiştirdiği görülmüştür. Özellikle Karar Ağacı(DTC) algoritmasında f-skor değeri önemli ölçüde iyileşmiştir. Tüm bunlara karşın Çapsal Tabanlı Fonksiyonlar(RBF) yönteminde ise sınıfların dengelemesi olumsuz sonuç doğurmuştur. Destek Vektör Makinesi(SVM) ve Rastgele Orman algoritmalarında ise çok düşük olan başarının değişmediği gözlenmiştir. Sınıf sayısının çok ve dengesiz sınıf dağılımlarının olduğu veri setlerinde Naive Bayes (NBC) yöntemi gayet iyi sonuç vermiştir. Sınıf ağırlıkları dengelenmesi neticesinde ise Karar Ağaçları (DTC) algoritmasında yüksek performans elde edilmiştir. Destek Vektör Makinesi(SVM) ve Rasgele Orman algoritmalarının ise çok sınıflı veri setlerine uygun olmadığı görülmüştür.

Received 17 August 2021; Received in revised form 25 August 2021; Accepted 21 September 2021;
Available online 25 September 2021;
doi: 10.46291/ICONTECHvol5iss3pp31-37

5. KAYNAKLAR

Agrawal, R., & Batra, M. 2013. A detailed study on text mining techniques. International Journal of Soft Computing and Engineering, 2(6), 118-121.

Berry, M. W. 2004. Survey of text mining. Computing Reviews, 45(9), 548.

Duygu Analizi. In International Artificial Intelligence and Data Processing Symposium (IDAP'16), September (pp. 17-18).

Jusoh, S., & Alfawareh, H. M. 2012. Techniques, applications and challenging issue in text mining. International Journal of Computer Science Issues (IJCSI), 9(6), 431.

Kaynar, O., Görmez, Y., Yıldız, M., & Albayrak, A. 2016. Makine öğrenmesi yöntemleri ile Schnabl, E., & Zenker, A. 2013. Statistical classification of knowledge-intensive business services (KIBS) with NACE Rev. 2. Karlsruhe: Fraunhofer ISI.

Nace. 2008. Konu: Avrupa Topluluğunda Ekonomik Faaliyetlerin İstatistiki Sınıflaması. https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_CLS_DLD_NOHDR&StrNom=NACE_REV2&StrLanguageCode=TR .Erişim:Ağustos, 2021.

Tobback, E., Naudts, H., Daelemans, W., de Fortuny, E. J., & Martens, D. 2018. Belgian economic policy uncertainty index: Improvement through text mining. International journal of forecasting, 34(2), 355-365.

Uyumsoft, 2020. Konu: Nace Kodlama Sistemi. Konu: https://www.uyumsoft.com/nace-kodu-nedir-ne-ise-yarar/ . Erişim: Ağustos, 2021

Van den Brakel, J. 2010. Sampling and estimation techniques for the implementation of new classification systems: the change-over from NACE Rev. 1.1 to NACE Rev. 2 in business surveys. In Survey Research Methods (Vol. 4, No. 2, pp. 103-119).

Zulfikar, W. B., Irfan, M., Alam, C. N., & Indra, M. 2017.. The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter. In 2017 5th International Conference on Cyber and IT Service Management (CITSM) (pp. 1-5). IEEE.

Received 17 August 2021; Received in revised form 25 August 2021; Accepted 21 September 2021;
Available online 25 September 2021;
doi: 10.46291/ICONTECHvol5iss3pp31-37