Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 25, Sayı 1, 112-120, 2021

Süleyman Demirel University
Journal of Natural andAppliedSciences
Volume 25, Issue 1, 112-120, 2021

DOI: 10.19113/sdufenbed.842460

Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması

Bilge ÖZLÜER BAŞER*[1], Metin YANGIN[2], E. Selin SARIDAŞ[3]

[1,2,3]Mimar Sinan Güzel Sanatlar Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, 34380, İstanbul, Türkiye

(Alınış / Received: 17.12.2020, Kabul / Accepted: 15.02.2021, Online Yayınlanma/ Published Online: 15.04.2021)

Anahtar Kelimeler
Diyabet,
Makine Öğrenmesi,
Sınıflandırma Algoritmaları

Özet: Diyabet, dünya çapında artan ve gerçekleşen ölümlerin önde gelen nedenlerinden biridir. Sürekli artan vaka sayısı diyabetin önlenmesi, erken teşhisi, tedavisi ve takibi konularında bilimsel çalışmalara ihtiyaç duyulduğunu göstermektedir. Son dönemlerde medikal alanda yaşanan teknolojik gelişmeler sayesinde elde edilen verinin analiz edilmesi, hastalıkların tanı ve tedavi sürecine olumlu katkılar yapmaktadır. Diyabet hastalığı kapsamında da araştırmacılar, hastalığın teşhis edilmesine yönelik, veriye dayalı sistematik yaklaşımlar geliştirmeye çalışmaktadırlar. Bu amaç doğrultusunda çalışmada, 1999-2008 yılları arasında ABD'de bulunan 130 hastanedeki 70000 kayda ait sağlık vakalarından elde edilmiş veri seti düzenlenerek, bireylerin diyabet durumuna göre sınıflandırılması hedeflenmiştir. Sınıflandırma için veri setine uygun makine öğrenmesi algoritmalarından yararlanılmış ve bu algoritmaların sonuçları performans ölçütlerine göre karşılaştırılmıştır. Elde edilen sonuçlara göre, en iyi performans gösteren beş sınıflandırma algoritması (Karar ağaçları, k-en yakın komşuluk, Lojistik regresyon, Naive Bayes ve Rastgele orman) değerlendirmeye alınmış olup en iyi doğru sınıflandırma performansı Rastgele orman algoritması ile elde edilmiştir.

Classification of Diabetes Mellitus with Machine Learning Techniques

Keywords
Diabetes Mellitus,
Machine Learning,
Classification Algorithms

Abstract: Diabetes is one of the leading causes of rising and occurring deaths worldwide. The ever-increasing number of cases indicates the need for scientific studies on the prevention, early diagnosis, treatment, and follow-up of diabetes. Analyzing the data obtained with the recent technological developments in the medical field makes positive contributions to the diagnosis and treatment process of diseases. As related to diabetes, researchers are trying to develop data-based systematic approaches to diagnose the disease. Following this purpose, the study aims to classify individuals according to their diabetes status by arranging a data set obtained from 70000 records of health cases in 130 hospitals in the USA between 1999-2008. Machine learning algorithms suitable for the data set are used for the classification and the results of these algorithms are compared regarding the performance criteria. According to the results, the best performing five classification algorithms (Decision trees, k-nearest neighborhood, Logistic regression, Naive Bayes, and Random forest) are evaluated and the best classification performance is obtained with the Random forest algorithm.

1. Giriş

Diyabet vücutta insülin hormonunun eksikliği, etkisizliği veya yeteri kadar üretilememesi sonucu gelişen, ayrıca kronik komplikasyonların karbonhidrat metabolizmasını bozan ve kandaki glikoz seviyesini arttıran bir hastalıktır. Yoğun susuzluk, yoğun açlık ve idrara sık çıkmak gibi semptomlarla görülen diyabet, tedavi edilmediği sürece hastada birçok komplikasyona neden olur. Zamanında önlem alınmadığı ve kan şekeri kontrol edilmediği takdirde özellikle damarlar üzerinde olumsuz etki göstermektedir. Şekerin toksik etkileri başta gözler, böbrekler, sinir uçları, kalp, beyin ve bacak damarları gibi pek çok organımızda ve dokumuzda kalıcı hasarlar oluşturabilmektedir [1]. Bu nedenle diyabette erken teşhis, birçok hasarın yaşanmaması için hayati önem teşkil etmektedir.

*İlgili yazar: bilge.baser@msgsu.edu.tr

Dünya Sağlık Örgütü'nün güncel verilerine göre, başta düşük ve orta gelirli ülkelerde çoğunlukla görülmek üzere dünyada yaklaşık 422 milyon insan diyabet hastasıdır ve her yıl gerçekleşen 1.6 milyon ölümün nedeni doğrudan diyabet ile ilişkilidir. Bu nedenle diyabet, dünyada gerçekleşen ölümlerin önde gelen nedenlerinden biri olarak kabul edilmektedir ve hem vaka sayısı hem de prevalansı sürekli olarak çarpıcı biçimde artmaktadır. Bu olumsuz tablo karşısında ülkeler, 2025 yılına kadar diyabetteki artışın global olarak durdurulması için hedef koymuş ve iş birliği yapmaya karar vermişlerdir [2].

Sağlık Bakanlığı'nın 2018 bütçe sunumunda yaptığı açıklamalara göre Türkiye, diyabet sıklığında OECD ülkeleri arasında ikinci sırada yer almaktadır ve yaklaşık 7 milyon diyabet hastasının var olduğu bilinmektedir. Türkiye Diyabet Vakfı'nın 2016 raporlarına göre ise, Türkiye'de diyabet tüm Avrupa ülkeleri arasında en hızlı artışı göstermektedir. Türkiye'nin bu tehlikeli konumu, diyabetin tanısı ve önlenmesi konusunda bilimsel çalışmalara ihtiyaç duyulduğunu vurgulamaktadır.

Teknolojik gelişmelerle birlikte yapay zekâ ve öğrenme teknikleri yardımıyla birçok hastalığın teşhis edilmesi mümkün hale gelmektedir. Bu sayede, hastalıkların teşhisi ve ilgili tetkiklerin raporlanması daha kısa sürede tamamlanmakta, bunun sonucunda da hastaların sağlık kuruluşunda harcadıkları süreyi azaltmaktadır. Günümüzde birçok ülkede akıllı hastane projelerine büyük yatırımlar yapılmaktadır. Bu uygulama hem sağlık kuruluşlarındaki yoğunluğu gidermekte hem de sistemi otomatize ederek gereken işgücü miktarını azaltmaktadır.

Çoğu araştırmacı, hastalıkların teşhisinde makine öğrenmesi algoritmaları kullanarak deneyler yürütmektedir. Makine öğrenmesi algoritmalarının tercih edilme nedeni ise, farklı hastalıkların teşhisinde daha az maliyetle, daha doğru ve daha hızlı sonuç vermeleridir. Çünkü, veri madenciliği ve makine öğrenmesi algoritmaları, çeşitli kaynaklardan gelen verileri birleştirip, büyük miktarda veriyi yönetebilme kabiliyeti sayesinde tahmin gücünü arttırmaktadır.

Son yıllarda medikal alanda giyilebilir teknolojiler de gelişimini hızla sürdürmektedir. Bu gelişmeler hem hastalığın ve tedavinin takibi bakımından hem de kişilerin hastalıkla mücadelesinde yaşam konforlarını arttırıcı etkiye sahip olmaları bakımından çok önemlidir. Diyabet alanında ise, kan testi yerine kişinin koluna takılan bir cihaz sayesinde anlık glikoz, glikoz trendi ve gittiği yön bilgisi küçük bir sensörden elde edilebilmektedir. Bu dijitalleşme ancak sensörlerden elde edilen verinin iyi analiz edilmesi ve doğru modellenmesi ile mümkündür.

Literatürde; "Pima Indians Diabetes" olarak bilinen veri seti [3] ile ilgili çalışmalar mevcut olup, bu çalışmalarda veri madenciliği ve sınıflandırma algoritmaları kullanılarak diyabet sınıfının tahmin edilmesi amaçlanmıştır. "Pima Indians Diabetes" veri setinin sıklıkla kullanılmasının nedeni Kuzey Amerika'da yaşayan Pima yerlilerinin genetik olarak diyabete yatkın olmaları ve diyabet görülme olasılığının yüksek olmasıdır. Diyabet hastalığının görülme sıklığının yüksek olduğu bu topluluğa ait olan veri seti, ABD Ulusal Diyabet ve Sindirim ve Böbrek Hastalıkları Enstitüsü (National Institute of Diabetes and Digestive and Kidney Diseases) tarafından oluşturulmuştur.

Joshi ve Shetty; bu veri seti üzerinde Bayesyen yaklaşım, Naive Bayes, J28, Rastgele orman, Rastgele ağaç, REP, k-NN, CART ve birleşmeli kural öğrenme algoritmalarını kullanarak performans karşılaştırmaları yapmışlardır [4]. Nidhi ve arkadaşları ise, sınıflandırma modeli oluşturmak için karar ağacı (J48), PART, Çok katmanlı algılayıcı ve Naive Bayes algoritmalarını kullanmışlardır [5].

"Pima Indians Diabetes" veri setini kullanarak diyabet hastalığı için sınıflandırma algoritmalarını inceleyen bir diğer çalışma da 2012 yılında Karegowda ve arkadaşları tarafından gerçekleştirilmiştir [6]. Bu çalışmada karar ağacı C4.5 ve k-ortalamalı kümeleme yöntemleri birleştirilerek karma bir model (hybrid model) oluşturulmuştur. İki aşamalı olarak çalıştırılan karma modelin doğru sınıflandırma oranı, yalnızca karar ağacı C4.5 yöntemi kullanılarak elde edilen sınıflandırma oranından daha yüksek bulunmuştur. Chen ve Pan ise, Wenzhou Medical Üniversitesi'nin ilk yardım hastanesi tarafından oluşturulan iki adet veri seti ile diyabet sınıfını tahmin etmeye çalışmışlardır. 35669 hasta üzerine yapılan klinik test sonuçlarına Adaboost.M1 ve LogitBoost olmak üzere iki ayrı makine sınıflandırması tekniğini uygulamışlardır. İki tekniğin de diyabet hastalığını sınıflandırma başarısı benzer bulunmuştur [7].

Literatürde "Pima Indians Diabetes" ile yapılan çalışmalarda yüksek doğruluk oranlarıyla sınıflandırma gerçekleştirilmiş olsa da veri setindeki gözlemlerin Pima yerlilerinden en az 21 yaşında olan kadınlardan oluşması, yani sınırlı bir gruba ait olması, elde edilen sonuçların genelleştirilebilmesi bakımından tartışmaya açıktır. Bu nedenle, literatürdeki çalışmalardan farklı olarak çalışmamızda, 1999-2008 yılları arasında ABD'de bulunan (Cerner Corporation, Kansas City, MO) 130 hastanedeki 70000 hastaya ait sağlık vaka verilerinden elde edilmiş veri seti kullanılmıştır [8]. Strack ve arkadaşları, bu veri seti ile çok değişkenli lojistik regresyon analizi çalışmış ve HbA1c ölçümü ile hastaneye yeniden yatış oranı arasındaki ilişkiyi modellemiştir [9] ancak veriye ilişkin yapılmış olan sınıflandırma çalışmasına rastlanmamıştır. Hem "Pima Indians Diabetes" veri setine göre daha kapsamlı hasta profili içermesi hem de açık kaynak olması, veri setinin tercih edilme nedenleridir.

Tablo 1. Çalışmada kullanılan değişkenler

Değişken adı	Türü	Açıklama ve değerler
Irk	Nominal	Değerler: Kafkasyalı, Asyalı, Afro-amerikalı, İspanyol ve diğer
Cinsiyet	Nominal	Değerler: Erkek, Kadın
Yaş	Nominal	Değerler: 10 yıl aralıklarla gruplanmış [0-10), [10-20), …, [90-100)
Ağırlık	Nominal	Değerler: [0-25), [25-50), ..., [175-200), [200- …]
Hastanede geçirilen süre	Nümerik	Hastaneye kabul ile taburcu edilmesi arasında geçen süre(gün)
Laboratuvar test sayısı	Nümerik	Hastanede gerçekleştirilen laboratuar testlerinin sayısı
Süreç sayısı	Nümerik	Test sayısı dışındaki süreç sayısı
İlaç sayısı	Nümerik	Kullanılan farklı ilaç sayısı
Poliklinik ziyaret sayısı	Nümerik	Hastanın bir yılda polikliniği ziyaret etme sayısı
Acil bölümü ziyaret sayısı	Nümerik	Hastanın bir yılda acili ziyaret etme sayısı
Yatarak tedavi edilme sayısı	Nümerik	Hastanın bir yılda yatarak tedavi edilme sayısı
Tanı sayısı	Nümerik	Sisteme girilen tanı sayısı
A1C test sonucu	Nominal	Değerler: Test sonucu> 8, 7<Test sonucu<8, Test sonucu<7 ve Test sonucu yok
Diyabetik ilaç geçmişi	Nominal	Evet, Hayır
Hastanın ilaç öyküsüne ilişkin değişkenler	Nominal	Sekiz ilaç (Metformin, Glimepiride, Glipzide, Glyburide, Pioglitazone, Rosiglitazone, Acarbose, İnsülin.) için dozaj durumu: Arttırma, azaltma, sabit ve ilaç verilmemesi
Sınıf değişkeni	Nominal	Diyabet durumu: Diyabetli, Diyabetli değil

2. Materyal ve Metot

2.1. Veri setinin düzenlenmesi

Çalışmada incelenen veri seti 55 özelliğe (nitelik veya değişken) sahip olup, ön-işlenmesi aşamasında "hasta numarası", "karşılama kimlik numarası", "ödeme kodu", "kabul yeri" değişkenleri, diyabet sınıflandırmasında bilgi verici özelliğe sahip olmamaları nedeniyle çalışma kapsamı dışında bırakılmıştır. Bu değişkenlerin yanı sıra, kayıp gözlemleri analize dahil edilmelerine olanak sağlayamayacak derecede fazla olan "başvurduğu medikal bölüm" değişkeni de veri setinden çıkarılmıştır. Ayrıca, dengesiz kategori yapısında olan, başka bir ifadeyle ilgili değişkene ait gözlemlerin yüzde olarak büyük çoğunluğunun tek bir kategoride toplandığı değişkenler (repaglinide, nateglinide vb.) sınıflandırma bakımından bilgi taşımadığı için çalışmaya dahil edilmemiştir.

Veri setinin analizi kapsamında kullanılan sınıf değişkeni daha önceki çalışmalardan farklı şekilde ele alınmıştır. Bunun için veri setinde var olan üç adet tanı kodları değişkenlerinden yararlanılmıştır. Verinin ham halinde her bir hastanın hastaneye ilk gelişi, ikinci gelişi, üç ve daha fazla sayıda gelişi süreçlerinde konulan hastalık teşhisleri, ICD-10 standartlarına uygun olarak kodlanmıştır. Bu üç değişkenin verdiği bilgiler kullanılarak ilgili hastaya ait üç farklı teşhis durumundan en az birinde diyabet tanısı konulması halinde, hastanın diyabet hastalığına sahip olduğu yönünde sınıf değişkeni ataması yapılmıştır.

Veri setinin düzenlenmesi aşamasında bir başka dikkate alınan konu ise, hastaların vücut ağırlığı bilgileridir. Verinin toplanması aşamasında ne yazık ki, birçok hastanın vücut ağırlığı bilgisi kayıt altına alınmamıştır. Ancak, bu konuda uzman kişilerin görüşlerine başvurulduğunda, vücut ağırlığının diyabet ile ilişkili olabileceği bilgisi elde edilmiştir. Bu nedenle incelenecek veri seti, ağırlık değişkenine ilişkin kaydın tutulmuş olduğu gözlemler üzerinden oluşturulmuştur. Vücut ağırlığı bilinen 3197 hasta dikkate alınarak veri temizleme işlemi yapılmıştır. Veri temizleme işleminde tekrarlı gözlemler söz konusu olduğu için hastaya atanmış olan hasta numarası esas alınarak ilgili hastaya ait sınıf değişkeni aynı olan gözlemler içerisinden sistemdeki son gözlem tutulup öncekiler çıkarılmıştır. Aynı hasta numarasına ait farklı sınıf değişkenine sahip gözlemler olması halinde ise, en az bir kez diyabet tanısı konulmuş olduğu için sınıfı diyabet olan gözlemler veri setinde tutulup olmayanlar devre dışı bırakılmıştır.

Veri temizleme işleminden sonra veri seti; 22'si bağımsız, biri de sınıf değişkeni olmak üzere 23 değişken ve 2705 gözleme indirgenmiştir.

2.2. Sınıflandırma modeli

Veri madenciliğinin temel yöntemlerinden biri olan sınıflandırma yöntemi, öğrenme algoritmasına dayanmaktadır. Büyük ölçekli veri setinde gizli olan örüntü (pattern) yapısını keşfetmek için kullanılır. Örüntü kavramı veri içerisinde gözlemlenebilir, ölçülebilir ve tekrar edilebilir bir bilgidir. Sınıflandırma algoritmaları ulaşılmak istenen bilgiyi hedefleyerek, veriyi ortak özelliklerine göre belirli gruplara (sınıflara) ayırmaktadır [10, 11].

Uygulamada sınıflandırma algoritmaları iki aşamalı olarak çalıştırılmaktadır. İlk olarak "eğitim verisi" (training data) olarak belirlenen veri setinin analiz

edilmesi ile sınıflandırma modeli oluşturulmaktadır. İkinci aşamada ise elde edilen sınıflandırma modeli, yeni bir veri kümesine uygulanarak belirlenen sınıfların veri içerisindeki varlığı araştırılmaktadır. Sınıf etiketlerinin tahmin edilmeye çalışıldığı ve modelin tahmin sonuçlarının performansının değerlendirildiği yeni veri seti "test verisi" (test data) olarak adlandırılmaktadır. Veri kümesini eğitim ve test verisi olarak ayrıştırma işlemi farklı şekillerde yapılabilir. Örneğin, veri setinin %60'lık kısmının eğitim, %40'lık kısmının test verisi olarak ayrıldığı, eğitim ve test kümelerinin rastgele atandığı ya da katlanarak değiştirildiği yöntemler mevcuttur. Bu çalışma kapsamında k-katlamalı çapraz doğrulama yöntemi kullanılmıştır.

2.2.1. k-Katlamalı çapraz doğrulama

k-katlamalı çapraz doğrulama yöntemi modelin değerlendirilmesi için önemli olup, modelin geliştirilme aşamasında aşırı öğrenmeyi ve eksik öğrenmeyi tespit etmekte [12] ve modelin test edilme aşamasında en iyi modeli oluşturmayı hedeflemektedir [13]. Aşırı öğrenme, modelin eğitim kümesindeki örüntüler yerine gözlemleri öğrenmesi durumunda ortaya çıkmaktadır. Bu durumda oluşturulan model eğitim aşamasında kullanılan veri kümesini öğrenir, ancak yeni gelen gözlemler için başarılı bir tahmin yapamaz. Genelde aşırı öğrenme modelleri eğitim aşamasını küçük hata oranı ile tamamlarken, test aşamasında büyük bir hata oranı ile tahmin etmektedir [13]. Eksik öğrenme ise, modelin gözlemlerdeki örüntüyü eksik bir şekilde öğrenmesi durumunda ortaya çıkmaktadır. Sınıflandırma modellerinde aşırı öğrenmeyi ve eksik öğrenmeyi önlemek için k-katlamalı çapraz doğrulama yöntemi kullanılmaktadır.

k-katlamalı çapraz doğrulama yönteminde, ilk olarak eğitim sürecinde kullanılacak eğitim kümesi karıştırılır ve eşit büyüklükteki k adet alt kümelere bölünür. Bu işlemler k–kez tekrarlanarak her iterasyonda sıradaki alt küme eğitim veri kümesinden çıkarılır ve test kümesi olarak kullanılır. Tüm parçalar için değerlendirme süreci tamamlandığında, çapraz doğrulama modeli tüm veriler için bir performans ölçütü ve sonuçlar üretmektedir [14].

Sınıflandırma modelinin başarısı doğru sınıfa atanan örnek sayısı ve yanlış sınıfa atanan örnek sayısı karşılaştırılarak belirlenmektedir. Model başarısını değerlendirmek için kullanılan temel kavramlar doğruluk, duyarlılık, özgüllük, hassasiyet ve AUC ölçütüdür. Bu ölçütleri hesaplayabilmek için Tablo 2. de verilmiş olan karışıklık matrisinden (confusion matrix) yararlanılır. Karışıklık matrisinde satırlar modelin öngördüğü sınıf sayılarını, sütunlar ise test kümesindeki gerçek sınıf sayılarını ifade etmektedir.

Tablo 2. Karışıklık matrisi

		Gerçek Değer	
		Doğru	**Yanlış**
Tahmin Edilen Değer	**Doğru**	Doğru Pozitif (DP)	Yanlış Pozitif (YP)
	Yanlış	Yanlış Negatif (YN)	Doğru Negatif (DN)

Karışıklık matrisindeki bu değerler kullanılarak sınıflandırma modelinin performansı ölçülmektedir. DP: gerçekte doğru olan değerin doğru tahmin edilmesi, YP: gerçekte yanlış olan değerin doğru tahmin edilmesi, YN: gerçekte doğru olan değerin yanlış tahmin edilmesi, DN: gerçekte yanlış olan değerin yanlış tahmin edilmesini ifade etmektedir.

2.2.2. Model performans değerlendirme ölçütleri

Karışıklık matrisindeki değerler kullanılarak sınıflandırma modelinin performansı ölçülmektedir. Bu çalışmada; doğru sınıflandırma oranı, duyarlılık, özgüllük, hassasiyet ve AUC (eğri altında kalan alan) performans ölçütleri dikkate alınmıştır.

Doğru sınıflandırma oranı (DSO)

Doğru sınıflandırılmış örnek sayısının toplam örnek sayısına oranıdır.

$$DSO = \frac{DP + DN}{N} \quad (1)$$

Duyarlılık (Sensitivity)

Doğru sınıflandırılmış pozitif örnek sayısının toplam pozitif örnek sayısına oranıdır. Gerçek pozitif oranı olarak da adlandırılır.

$$Duyarlılık = \frac{DP}{DP + YN} \quad (2)$$

Özgüllük (Specificity)

Doğru sınıflandırılmış negatif örnek sayısının toplam negatif örnek sayısına oranıdır. Gerçek negatif oranı olarak da adlandırılır.

$$\text{Özgüllük} = \frac{DN}{YP + DN} \quad (3)$$

Hassasiyet (Precision)

Doğru pozitif değerleri tahmin etme oranıdır.

$$Hassasiyet = \frac{DP}{DP + YP} \quad (4)$$

AUC (Eğri altında kalan alan)

AUC (Area Under Curve) değeri sınıflandırma modelinin sınıfları ayırt edebilme başarısını göstermekte ve Alıcı İşletim Karakteristiği (AİK) eğrisi altında kalan alanı ifade etmektedir. AİK eğrisi

(0,0) ile (1,1) arasında artan bir fonksiyondur. AİK puanı olarak da adlandırılan AUC değeri ne kadar yüksek ise modelin sınıflandırma başarısı da o kadar artmaktadır. AUC, en büyük "1" ve en küçük "0.5" değerini alabilmektedir. AUC değerinin 0.5 olması sınıflandırma modelinin başarısız olduğunu ve sınıf atamalarının rastgele yapıldığını göstermektedir.

2.3. Sınıflandırma algoritmaları

Çalışma kapsamında ele alınan sınıflandırma problemi için literatürde var olan makine öğrenmesi sınıflandırma algoritmalarından yararlanılarak bu algoritmalar içerisinden en iyi performansı veren Naive Bayes, Rastgele orman, Karar ağaçları, Lojistik regresyon, k-En yakın komşuluk algoritmaları ele alınmıştır.

2.3.1. Naive Bayes

Bayes sınıflandırıcıları istatistiksel sınıflandırıcı ailesine ait olmakla beraber, sınıf üyelik olasılıkları yardımıyla belirli bir sınıfa ait olma olasılığını tahmin ederek atama yapmaktadır. Bu sınıflandırma algoritmasının temeli, Bayes teoremine dayanmaktadır ve diğer öğrenme yöntemlerinden farklı olarak eğitim verisinin çeşitli kombinasyonlarına ilişkin sıklıkları hesaplamasıdır [15]. Naive Bayes'in avantajı, kolay uygulanması ve genellikle iyi sonuç vermesi iken, dezavantajı ise sınıf ve nitelik bakımından bağımsızlık varsayımına ihtiyaç duymasıdır. Naive Bayes algoritması olayların birbirinden bağımsız olduğu varsayımı altında kullanılmakta ve sonsal olasılık değerinin hesaplanması yardımı ile çalışmaktadır.

2.3.2.Rastgele orman

Rastgele orman, içerisinde birçok sınıflandırma ağacı bulundurmakla beraber, doğru sınıflandırma değerini çok fazla arttıran bir yöntemdir [16]. Yeni bir örneğin sınıflandırılmasında izlenen adımlarda, incelenen örneğe ait girdi vektörünün, ormandaki her bir ağaç tarafından tek tek sınıflandırılması olayına ağaç oylaması denmektedir [17].

Rastgele orman algoritması, girdi değişkeninin çok olduğu büyük verilerde iyi performans göstermesinin yanı sıra eksik verilerde de yüksek tahminler gerçekleştirmesi bakımından büyük kolaylık sağlamaktadır.

Aynı zamanda rastgele orman algoritması, model geliştirilirken örnekleme boyutunun parametresini belirleyerek her sınıftan ne kadar örnekleme yapılacağını, yerine koyma yöntemi ile belirlemektedir. Örnekleme boyutu parametresi, rastgelelik etkisini sağlayarak, ormandaki her ağaç verisinin farklı yönlerini görmektedir. Bu parametre sayesinde, dengesiz sınıf dağılımına sahip veri kümelerinde model geliştirilirken, dengesizlik problemini aşmaya ve modelin performansını arttırmaya da yardımcı olmaktadır.

2.3.3. Karar ağaçları

Karar ağaçları sınıflandırma problemlerinde yaygın olarak kullanılan algoritmalardan biridir. Veri madenciliğinde karar ağaçları, sınıflandırma ve regresyon ağaçlarını göstermek için kullanılmaktadır. Karar ağaçlarının avantajı, oluşturulmasının ve yorumlanmasının kolay olmasıdır.

Karar ağaçları, denetimli öğrenme için kullanılan dağılımlardan bağımsız bir öğrenme yöntemi türüdür. Yapısı, aynı bir ağaç gibi kök düğüm, dal ve yapraklardan oluşmaktadır. Yaprak kısmında oluşan değer, çıktı olarak adlandırılmaktadır ve araştırılan problem, sınıflandırma problemi ise, sınıf etiketi; regresyon problemi ise, sayısal bir değeri almaktadır. Karar ağacı, kullanılan verinin durumundan etkilenmekte, eğer veri seti karmaşık ise ağaç dallanıp büyümektedir [18, 19]. Karar ağacı, sınıflama, özellik ve hedefe göre karar düğümleri (decision nodes) ve yaprak düğümlerinden (leaf nodes) oluşan ağaç yapısı formunda bir model oluşturan bir sınıflandırma yöntemidir. Özellik veya karar düğümü seçim ölçüsü, verilen eğitim başlıklarını tanımlayan her özellik için bir sıralama sağlamakta ve hangi özelliğin seçileceğine karar vermektedir. Özellik seçimi yani hangi düğümün seçileceğinin belirlenmesinde kullanılan ölçüler: kazanım değeri (information gain), kazanım oranı (gain ratio) ve gini indeksi (gini index) olarak sıralanabilir. Bir veri kümesinden birden fazla ağaç elde edilmesine rağmen, en küçük boyutlu ağaç tercih edilmektedir. Değişken seçimi sırasında algoritmanın karar ağacı modelindeki döngüden çıkabilmesi için bulunduğu düğümdeki bütün elemanların aynı sınıfa atanmış olması gerekmektedir. Bu durumda yapraklarda bütün elemanlar aynı sınıfta yer alacağından ve sınıflandırma yapılacak değer kalmayacağından, karar ağacı modelindeki döngü durdurulur ve karar ağacı modeli oluşumu tamamlanmış olur. Kısacası, benzerlik göstermiş elemanların sınıflara dağılımı yapılmış olur.

2.3.4. Lojistik regresyon

Lojistik regresyon analizi, ele alınan veri setindeki gözlemlerin gruplara atanmasında kullanılan yöntemlerden birisidir. Sınıf sayısı bilinen lojistik regresyon analizinde var olan veriler kullanılarak sınıflandırma modeli elde edilir ve elde edilen bu model sayesinde veriye eklenecek yeni gözlemlerin sınıflara atanması sağlanabilmektedir [18, 19].

Aynı zamanda lojistik regresyon, bağımlı değişkenin yani sınıf değişkeninin iki veya daha fazla kategorili olduğu durumlarda bağımsız değişkenlerle olan neden-sonuç ilişkisini belirlemede kullanılan bir yöntemdir. Lojistik regresyon analizinde bağımsız değişkenlerle bağımlı değişken arasındaki ilişkinin

önemli olup olmadığı incelenerek, ilgilenilen değişkenin modelde var olup olmadığı durumlar için elde edilen tahmin değerleri ile gözlenen değerlerin karşılaştırılması yapılmaktadır. İlgilenilen değişkenin modelde yer aldığı durumda daha iyi, daha doğru tahminler elde edilmesi; o değişkenin model için önemli bir değişken olduğu şeklinde yorumlanır. Araştırmacıların, çalıştıkları konuda birden çok etkenin olması halinde, etkenlerin bağımlı değişken üzerine etkisini tek tek öğrenmenin yanı sıra, bunların birlikte bağımlı değişken üzerindeki etkisini de bilmek ve incelemek istemeleri durumunda tercih ettikleri yöntemlerden biridir.

Lojistik regresyon modellerinin özellikle tıp alanındaki uygulamalarında bağımsız değişkenler; risk değişkenleri ya da bir hastalığın ortaya çıkıp çıkmamasını belirleyen değişkenlerdir. Bu değişkenlerin tespiti, erken tanı ve hastalığa neden olan etkenlerle mücadelede önemli bir yer tutmaktadır.

2.3.5. *k*-En Yakın Komşuluk

Bu algoritmanın çalışma prensibi ilk olarak veri kümesini eğitim ve test verisi olmak üzere ikiye bölmektir. Test kümesinden bir gözlemin sınıflandırılmasında nitelik uzayında ele alınan gözlemin, eğitim kümesindeki gözlemlerin her birine olan uzaklığı ayrı ayrı hesaplanmaktadır. Burada sınıflandırma sırasında çıkarılan özelliklerden, sınıflandırılmak istenen yeni gözlemin daha önceki gözlemlerden *k* tanesine olan yakınlığına bakılmaktadır.

Önceden belirlenen bir *k* sayısı kadar en yakın komşularının hangi sınıflara dâhil olduklarına bakılarak, en çok gözlem sayısı hangi sınıfta ise, ele alınan örnek veri o sınıfa dâhil edilir. Yani yeni gelen bir örneğin, en yakın *k* adet komşusundan çoğunluğunun üyesi olduğu sınıfa dâhil edilmesi anlamına gelir. Eğer *k* sayısı 1 olursa bu durumda en yakın komşusunun sınıfına dâhil edilmektedir. Bu algoritmada k sayısının seçimi, sonucu belirlemede kritik bir öneme sahiptir. *k* sayısının çift olması, gözlemin tüm sınıflara da yakın olmasına halinde eşitlik durumu yaratabileceği için, *k* genellikle tek sayı olarak belirlenmektedir. Yapılan çalışmalarda, en iyi sonuçların genellikle *k* sayısının 1, 3 ve 5 değerlerini aldığında elde edildiği görülmüştür [20]. Algoritmanın performansını *k* değerine ek olarak uzaklık hesaplama yöntemi de etkilemektedir. Bu nedenle, farklı uzaklık hesaplama yöntemleri kullanılarak elde edilen sonuçların karşılaştırmalı analizi yapılabilmektedir.

3. Bulgular

Bu çalışmada gerçekleştirilen analizler için WEKA paket programı ve R programlama dili kullanılmıştır. Elde edilen veri setinde hastanın diyabet hastalığına yakalanması durumu incelenen sınıf olarak belirlenmiştir. Ele alınan veri setinde 727 hastanın diyabet, 1978 hastanın ise diyabet olmadığı tespit edilmiştir. Bu şekilde sınıfların eşit olarak temsil edilmediği veri kümesi, dengesiz bir veri kümesi olarak değerlendirilir. Makine öğrenimi algoritmaları genellikle performans ölçütlerinden doğru sınıflandırma oranını kullanarak değerlendirme yapmaktadır ancak, veriler dengesiz olduğunda bu yaklaşım uygun olmamaktadır [21].

Verilerin yüksek boyutlu olmasının birçok sınıflandırıcı için sınıf dengesizliği sorununu arttırdığı görülmüştür [22]. Yüksek boyutluluk, her bir sınıflandırıcı türünü farklı bir şekilde etkilemektedir. Genel kanı, eğitim verileri ile gerçek değerler arasındaki büyük tutarsızlıkların, daha büyük bir örnekleme değişkenliği olan azınlık sınıfında ortaya çıkma olasılığının daha yüksek olduğudur [23].

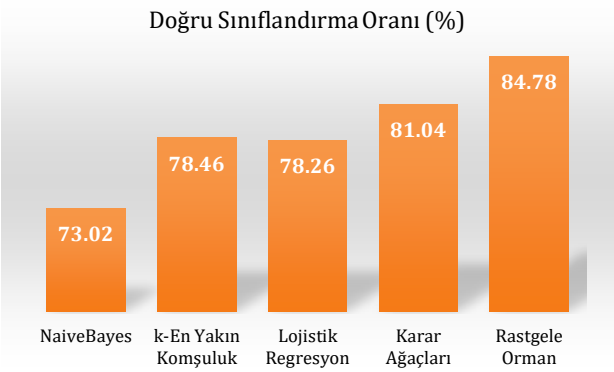
Bu sorunların giderilmesi amacıyla, azınlık sınıfının sentetik örnekler yaratılarak örneklendiği bir aşırı örnekleme yaklaşımı olan "Azınlık Aşırı Örnekleme Tekniği (SMOTE)" önerilmiştir [21]. Potansiyel olarak basit örneklemeden daha iyi performans göstermekte ve yaygın olarak kullanılmaktadır [23]. Azınlık sınıfı, her bir azınlık sınıfı örneği alınarak ve en yakın komşuluk sınıflarının herhangi birine/hepsine katılan çizgi parçaları boyunca sentetik örnekler vererek aşırı örneklenmektedir [21].

Çalışmada söz konusu olan dengesiz sınıf değişkeni (diyabet olma) için SMOTE yöntemi kullanılarak üretilen sentetik verilerle beraber diyabet sınıfı 1977 ve diyabet olmayan sınıf 1978 olarak elde edilerek, bir başka ifadeyle dengeli hale getirilerek analize dahil edilmiştir.

Çalışma kapsamında incelenen sınıflandırma problemi için oluşturulan karışıklık matrisinde;

DP: Veri setinde, gerçekte diyabet olan hastanın model tarafından diyabet olarak tahmin edilmesine,
DN: Veri setinde, gerçekte diyabet olmayan hastanın model tarafından diyabet olmayan olarak tahmin edilmesine,
YP: Veri setinde, gerçekte diyabet olmayan hastanın model tarafından diyabet olarak tahmin edilmesine,
YN: Veri setinde, gerçekte diyabet olan hastanın model tarafından diyabet olmayan olarak tahmin edilmesine karşılık gelmektedir.

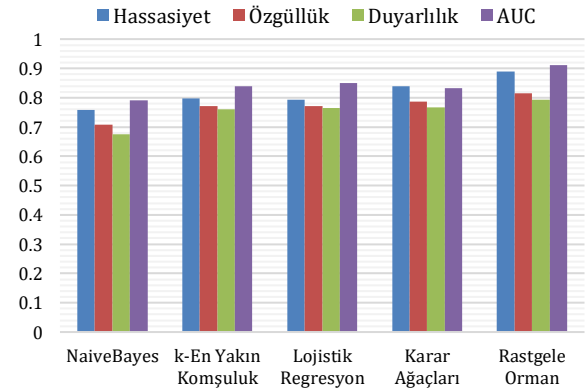
Çalışmada kullanılan algoritmalardan en yüksek performans gösteren beş algoritmanın doğru sınıflandırma oranlarının grafiksel gösterimi Şekil 1'de görülmektedir. Buna göre, rastgele orman % 84.78 ile diğerlerinden daha yüksek oranda doğru sınıflandırma yapmıştır.



Şekil 1.Sınıflandırma algoritmalarının doğru sınıflandırma oranlarının karşılaştırılması

Söz konusu algoritmaların hassasiyet, özgüllük, duyarlılık ve AUC gibi diğer model performans ölçütleri karşılaştırıldığında Şekil 2'de verilmiş olan grafiksel yapıya ulaşılmaktadır. Burada hassasiyet ölçütü diyabetli olduğu tahmin edilen hastaların gerçekte de diyabet hastası olmasının etkililiğini göstermektedir. Özgüllük ölçütü ise, diyabet hastası olmayan olarak tahmin edilen hastaların gerçekte de diyabet hastası olmayan toplam örnek sayısına oranını temsil etmektedir. Duyarlılık ölçütü, diyabet hastası olarak tahmin edilen hastaların gerçekte diyabetli hastaların toplam sayısına oranını verirken, AUC ölçütü de diyabet olan ve diyabet olmayan sınıfları ayırt etme başarısını göstermektedir.

Şekil 2 dikkate alındığında, duyarlılık ölçütüne göre tüm algoritmalar birbirine yakın performans gösterseler de diğer ölçütler dikkate alındığında rastgele orman algoritması daha yüksek performansıyla öne çıkmaktadır.



Şekil 2. Sınıflandırma algoritmalarının performanslarının karşılaştırılması

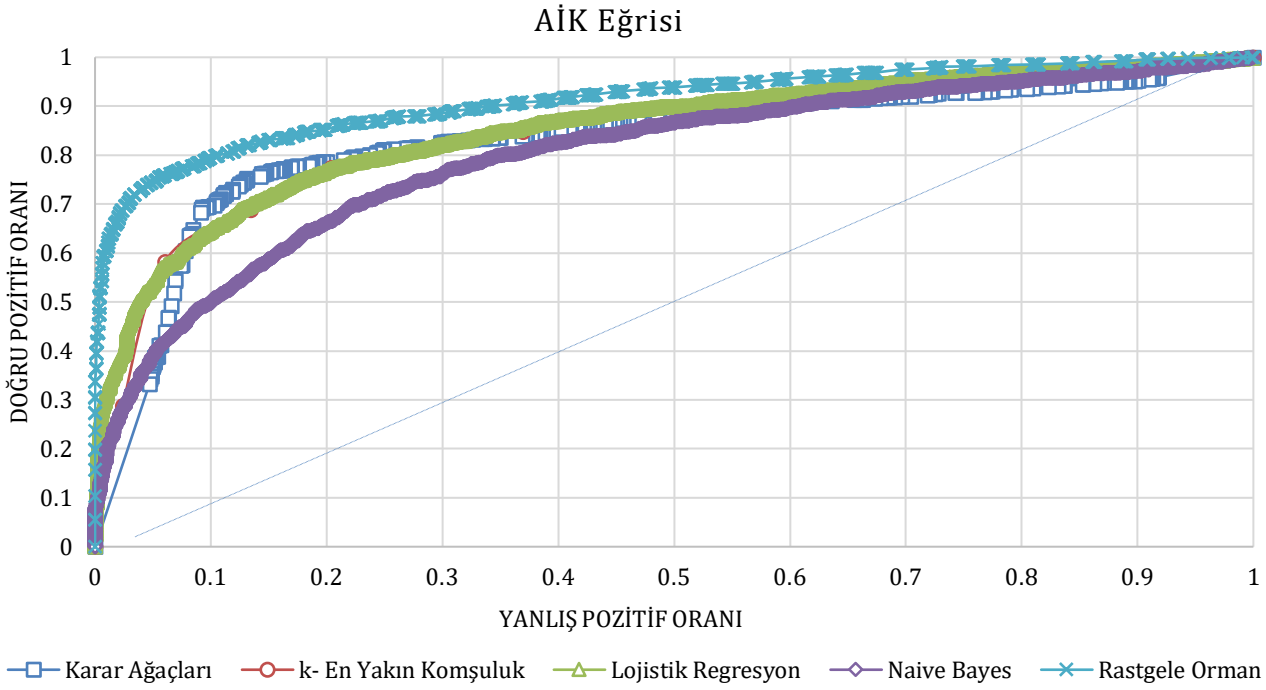
Sınıflandırma modelinin sınıfları ayırt edebilme başarısını gösteren AUC değerleri Şekil 3'te verilmiş olan grafik üzerinde incelendiğinde, ilgilenilen tüm algoritmaların 0.5'ten yüksek değere ulaşabildiği ve sınıf atamalarının rastgele yapılmadığı görülmektedir. Bununla birlikte, rastgele orman algoritmasının AUC ölçütü bakımından en iyi performansı sağladığı; karar ağaçları, lojistik regresyon ve k en yakın komşuluk algoritmalarının yaklaşık performans sergilediği, en düşük performans değerinin ise Naive Bayes algoritması ile elde edildiği görülmektedir.

Kullanılan algoritmaların performans ölçütlerinin sayısal değerleri Tablo 2'de görülmektedir. Buna göre, rastgele orman algoritması hassasiyet, özgüllük, duyarlılık ve AUC ölçütlerine göre sırasıyla (0.89, 0.814, 0.793, 0.912) değerleri ile en yüksek performansı göstermiştir.

Tablo 2.Farklı sınıflandırma algoritmalarına ilişkin performans ölçütleri

	Hassasiyet	***Özgüllük***	***Duyarlılık***	***AUC***
Naive Bayes	0.758	0.708	0.676	0.791
k-En Yakın Komşuluk	0.798	0.772	0.761	0.839
Lojistik Regresyon	0.793	0.772	0.764	0.850
Karar Ağaçları	0.840	0.786	0.767	0.832
Rastgele Orman	**0.890**	**0.814**	**0.793**	**0.912**

Dolayısıyla, çalışmada ele alınan veri seti için kişinin diyabet hastası olup olmadığının belirlenmesinde en iyi sınıflandırmayı kullanılan performans ölçütlerine göre, *Rastgele orman* algoritmasının sağladığı söylenebilir.



Şekil 3. Sınıflandırma Algoritmalarının AUC ölçütlerinin karşılaştırılması

4. Tartışma ve Sonuç

Bu alanda giderek artan büyük verinin analiz edilmesi için sistematik yaklaşımlara ihtiyaç duyulmaktadır. Bu nedenle, makine öğrenme tekniklerinin konu ile ilgili araştırmalarda uygulanması büyük önem kazanmış, literatürde sıklıkla karşılaşılır hale gelmiştir.

Konunun gelişime açık olması, çalışma kapsamında açık kaynak olarak referans gösterilen bir veri setinin seçilme nedenidir. Elde edilen sonuçlara göre, sağlık kuruluşuna gelen yeni bir hastanın çalışma kapsamında incelenen 22 bağımsız değişken bilgisi kullanılarak diyabet hastası olma olasılığı %84,78 doğruluk ve 0.912 AUC değeri ile tahmin edilebilecektir. Bu bilgi, hastalığın teşhisinde sağlık kuruluşunda harcanacak süreyi azaltıcı yönde etkiye sahip olacaktır ve dolayısıyla sağlık kuruluşlarında yaşanan yoğunluğu giderici yönde katkı sağlayacaktır. Buna ek olarak, sensörlerden elde edilen verilerin işlenmesi ile hastanın bazı tetkikleri yaptırmasına gereksinim duyulmayacaktır. Bu da yine zaman ve maliyet bakımından tasarruf sağlayıp, hastanın yaşam konforunu arttıracaktır. Tüm bu avantajların yanı sıra, elde edilen verinin hızlıca analiz edilmesi, hastalığın erken tanılanması ve tedavisine daha kısa sürede başlanması konularında büyük katkı sağlayacaktır.

Gelecek araştırmalarda, diyabet hastalığına ilişkin farklı veri setleri üzerinde sınıflandırma algoritmaları kullanılarak elde edilen sonuçların geliştirilmesi yönünde çalışılması planlanmaktadır.

Teşekkür

Bu çalışma, Mimar Sinan Güzel Sanatlar Üniversitesi, Bilimsel Araştırma Projeleri birimi tarafından 2019-30 numaralı proje ile maddi olarak desteklenmiştir.

Etik Beyanı

Bu çalışmada, "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması gerekli tüm kurallara uyulduğunu, bahsi geçen yönergenin "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbirinin gerçekleştirilmediğini taahhüt ederiz.

Kaynakça

[1] Siva, Z. O. http://www.diyabet.com/diyabet-hakkinda/diyabet-nedir/diyabet-nasil-bir-hastaliktir.html (Erişim Tarihi: 10.01.2020).

[2] Anonim, Dünya Sağlık Örgütü, "World Health Organization". https://www.who.int/health-topics/diabetes#tab=tab_1 (Erişim Tarihi: 05.06.2020).

[3] Kaggle, 2018. http://www.kaggle.com/kumargh/pimaindiansdiabetescsv (Erişim Tarihi: 10.01.2020).

[4] Joshi S., Priyanka Shetty, S. R. 2015. Performance Analysis of Different Classification Methods in Data Mining for Diabetes Dataset using WEKA Tool. International Journal on Recent and Innovation Trends in Computing and Communication, 3(3), 1168-1173.

[5] Walia N., Kumar M., Kakkar L. 2018. Classification of Diabetes Patient by using Data

Mining Techniques. International Journal for Research in Engineering Application & Management, 4(5), 347-351.

[6] Karegowda, A. G., Punya, V., Jayaram, M. A., Manjunath, A. S. 2012. Rule Based Classification for Diabetic Patients using Cascaded k-means and Decision Tree C4. 5. International Journal of Computer Applications, 45(12), 45-50.

[7] Chen, P., Pan, C. 2018. Diabetes Classification Model Based on Boosting Algorithms. BMC Bioinformatics, 19(1), 1-9.

[8] https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008# (Erişim Tarihi: 10.12.2019).

[9] Strack B., DeShazo J. P., Gennings C., Olmo J. L., Ventura S., Cios K. J., Clore J. N. 2014. Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records. BioMed Research International, Article ID 781670, 11s.

[10] Çınar, A. 2019. Veri Madenciliğinde Sınıflandırma Algoritmalarının Performans Değerlendirmesi ve R Dili ile Bir Uygulama. Öneri dergisi, 14(51), 90-111.

[11] Han, J., Kamber, M., Pei J. 2011. Data Mining: Concepts and Techniques. Third edition. The Morgan Kaufmann Series in Data Management Systems, 5(4), 83-124.

[12] Singh, A., Tiwari, V., Tentu, A. N. 2018. A Machine Vision Attack Model on Image Based CAPTCHAs Challenge: Large Scale Evaluation. In International Conference on Security, Privacy, and Applied Cryptography Engineering, Springer, Cham, December 15-19, Kanpur, India, 52-64.

[13] Arlot, S., Celisse, A. 2010. A Survey of Cross-validation Procedures for Model Selection. Statistics Surveys, 4, 40-79.

[14] Wiens, T. S., Dale, B. C., Boyce, M. S., Kershaw, G. P. 2008. Three Way k-fold Cross-validation of Resource Selection Functions. Ecological Modelling, 212(3-4), 244-255.

[15] Mitchell, M. T. 1997. Machine Learning. Sinagapore, TheMcGraw-Hill, 414s.

[16] Breiman, L. 2001. Random Forests. Machine Learning, 45(1), 5-32.

[17] Rokach, L., Maimon, O. Z. 2008. Data Mining with Decision Trees: Theory and Applications. 2nd Edition, World Scientific, 305s.

[18] Hosmer Jr., D. W., Lemeshow, S., Sturdivant, R. X. 2013. Applied Logistic Regression. 3rd Edition, John Wiley & Sons, 510s.

[19] Kuyucu, Y.E. 2012. Lojistik regresyon analizi (LRA), yapay sinir ağları (YSA) ve sınıflandırma ve regresyon ağaçları (CART) yöntemlerinin karşılaştırılması ve tıp alanında bir uygulama. Gaziosmanpaşa Üniversitesi, Sağlık Bilimleri Enstitüsü, Yüksek Lisans Tezi, 128s, Tokat.

[20] Dudoit, S., Fridlyand, J., Speed, T. P. 2002. Comparison of Discrimination Methods for the Classification of Tumors using Gene Expression Data. Journal of the American Statistical Association, 97(457), 77-87.

[21] Chawla, N. V., Bowyer, K. W., Hall, L. O., Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321-357.

[22] Blagus, R., Lusa, L. 2010. Class Prediction for High-dimensional Class-imbalanced Data. BMC Bioinformatics, 11(523), 1-17.

[23] Blagus, R., Lusa, L. 2013. Improved Shrunken Centroid Classifiers for High-dimensional Class-imbalanced Data. BMC Bioinformatics, 14(64), 1-13.