Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Apache Spark ile Makine Öğrenmesi Destekli Diyabet Rahatsızlığı Tahmini

Emre YILDIRIM [a,*], Ali ÇALHAN [b]

[a] *Bilgisayar Teknolojileri Bölümü, Osmaniye Meslek Yüksekokulu, Osmaniye Korkut Ata Üniversitesi, Osmaniye, TÜRKİYE*

[b] *Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, Düzce Üniversitesi, Düzce, TÜRKİYE*

** Sorumlu yazarın e-posta adresi: emreyildirim@osmaniye.edu.tr*

DOI: 10.29130/dubited.999048

Öz

Diyabet rahatsızlığı, insan vücudunun organlarını etkileyen kritik sağlık sorunlarından biridir. Bu nedenle, diyabet, 21. yüzyılda küresel bir sağlık sorunu olarak kabul edilmektedir. Bu rahatsızlığın sonucu olarak ortaya çıkan sorunlardan kaçınmak ve onları ağırlaşmadan önce tedavi etmek için diyabet rahatsızlığını tahmin edip işleyebilen bir sisteme ihtiyaç duyulmaktadır. Son yıllarda, sağlık alanında birçok rahatsızlığın erken teşhisi için çeşitli teknolojik araçlar ve uygulamalar kullanılmaktadır. Bu uygulamalardan birisi de veri madenciliği ve makine öğrenmesi teknikleri yardımıyla hastalığın erken teşhisi için analizlerin gerçekleştirilmesidir. Bu araştırmada, son zamanlarda büyük veri işlemede oldukça popüler olan Apache Spark teknolojisi ile diyabet rahatsızlığı analizleri gerçekleştirilmektedir. Aynı zamanda analizlerde tahmin için kullanılan Apache Spark MLlib kütüphanesindeki beş farklı makine öğrenmesi sınıflandırma algoritmalarının performansları karşılaştırılmış ve Rasgele Orman (RO) algoritmasının en iyi performansa sahip olduğu görülmektedir. Gerçekleştirilen analizler sonucunda kullanılan Apache Spark teknolojisinin bu tarz rahatsızlıkların belirlenmesinde kullanılabileceğini göstermektedir.

Anahtar Kelimeler: *Apache Spark, Diyabet Rahatsızlığı, Makine Öğrenmesi.*

Machine Learning Supported Diabetes Prediction with Apache Spark

Abstract

Diabetes is one of the critical health problems that affect the organs of the human body. Therefore, diabetes is recognized as a global health problem in the 21st century. To avoid the problems that arise as a result of the diabetes and to treat it before it worsen, there is a need for a system that can predict and process diabetes. In recent years, various technological tools and applications have been used for the early diagnosis of many diseases in the field of health. One of these applications is to perform analyzes for early diagnosis of the disease with the help of data mining and machine learning techniques. In this study, diabetes analyzes are carried out with Apache Spark technology, which has been very popular in big data processing recently. So, the performances of five different machine learning classification algorithms in the Apache Spark MLlib library used for prediction in the analysis are compared and it is seen that the Random Forest (RO) algorithm has the best performance. The results of the analyzes show that the Apache Spark technology used can be used to detect such health problems.

Keywords: *Apache Spark, Diabetes, Machine Learning*

Geliş: 22/09/2021, Düzeltme: 03/11/2021, Kabul: 22/11/2021

I. GİRİŞ

Günümüzde sağlık hizmeti sorunlarına çözüm sağlamada büyük veri ve bulut teknolojileri önemli bir role sahiptir. Sağlık hizmetleri ile ilgili veriler gün geçtikçe büyük oranda artmaktadır. Bu veriler, hastalıklar karşısında ölüm oranını düşürmek ve hastalıklara erken teşhis konabilmesi için makine öğrenmesi gibi çeşitli yöntemlerle analiz edilmektedir. En yaygın görülen kronik sağlık sorunlarından biri diyabet rahatsızlığıdır. Uzun vadede bu sorun, uygun olmayan ilaçların kullanılması durumunda şeker hastasının gözlerine, kalbine, böbreklerine ve sinirlerine zarar verebilmekte ve dahası ölümüne neden olabilmektedir. Dünya Sağlık Örgütü'ne göre, dünyada diyabet rahatsızlığına sahip 422 milyon kişi bulmakta ve bu hastalık nedeniyle her yıl 1.6 milyon kişi hayatını kaybetmektedir [1]. Diyabet rahatsızlığının teşhisinin konması karmaşıktır. Teşhisin, kesin ve yetkin bir şekilde yapılması kritik bir görevdir. Bu görev çoğunlukla uzman doktorun kararı üzerine yapılmaktadır. Bu durum, aşırı zaman ve maliyete neden olabilmektedir. Bunun yerine, güncel teknolojiler yardımıyla hastalıkla ilgili toplanan veriler karar destek sistemlerinden yararlanarak otomatik bir tıbbi sistem ile analiz edilebilmektedir. Bu yöntem, daha verimli bir şekilde diyabet rahatsızlığının teşhisinin konulmasına yardımcı olabilmektedir. Buna göre, diyabet rahatsızlığı ile ilgili verileri analiz edebilmek için farklı makine öğrenme yöntemleri ve veri madenciliği araçları kullanılmaktadır. Örneğin, Apache Flink [2], Apache Hadoop [3] ve Apache Spark [4] gibi büyük verileri rahat bir şekilde işleyebilen farklı büyük veri analizi teknolojileri bulunmasına rağmen literatürdeki çalışmalar incelendiğinde, bu platformların çok fazla kullanılmadığı görülmektedir [5, 6, 7, 8].

Han vd. [5] çalışmalarında, diyabet rahatsızlığının tahmini için naive bayes (NB), karar ağacı (KA) ve destek vektör makinesi (DVM) olmak üzere üç farklı sınıflandırma algoritmasını kullanmışlardır. Deneysel sonuçlar, NB algoritmasının %76,30'luk maksimum doğruluk ile en iyi performansı verdiğini göstermiştir.

Kumar ve Pranavi [6] diyabet rahatsızlığının tahmininde en iyi sınıflandırma algoritmasını belirlemek için doğruluk, kappa, kesinlik, duyarlılık ve özgünlük gibi çeşitli ölçütleri değerlendirmişlerdir. Çalışmada kullanılan sınıflandırma algoritmaları ise RO, DVM, k-En Yakın Komşu (k-NN), Sınıflandırma ve Regresyon Ağaçları (CART) ve Gizli Dirichlet Ayırımı (LDA) algoritmalarıdır. Elde edilen sonuçlara göre, RO algoritmasının verileri daha doğru ve hassas tahmin ettiğini göstermektedir.

Zou vd., [7] çalışmalarında, diyabet rahatsızlığını tahmin etmek için RO, KA ve yapay sinir ağı (YSA) sınıflandırıcılarını kullanmaktadırlar. Analizler sonucunda, RO %80'lik doğruluk oranı ile en iyi tahmini yapmaktadır.

Barakat vd. [8] diyabet rahatsızlığının teşhisi için DVM sınıflandırma algoritmasının kullanılmasını önermektedir. Analiz için kullanılan diyabet rahatsızlığı veri setinin sonuçları, DVM'nin, %94'lük doğruluğu, %93'lük kesinlik ve %94'lük özgünlük ile diyabet tahmini için uygun bir araç olarak kullanılabileceğini belirtmektedir.

Mir ve Dhage [9] NB, DVM, RO ve CART sınıflandırma algoritmalarını kullanarak diyabet hastalığını tahmin için WEKA aracını kullanmışlardır. Çalışmada, DVM algoritması %79,13 maksimum doğrulukla diyabet rahatsızlığını tahmin ettiği gözlemlenmiştir.

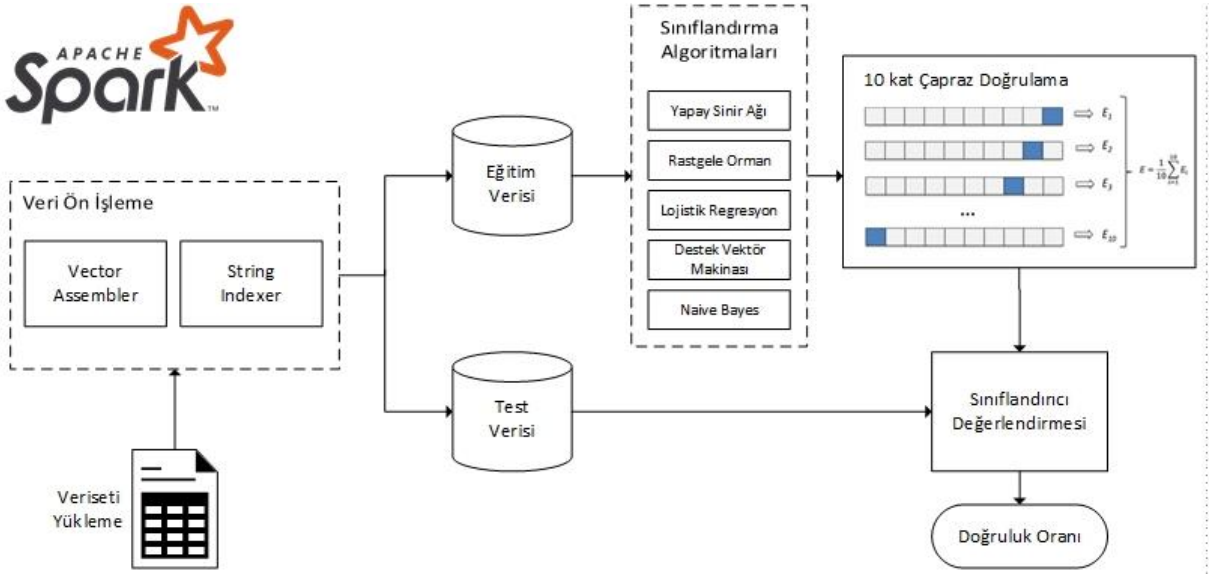
Hassan ve Shaheen [10], sosyal medya üzerinden gelen anlık veriler ile diyabet rahatsızlığını tahmin eden bir sistem önermektedirler. Sistemin amacı, diyabet tahmini için en iyi makine öğrenmesi modelini bulmaya çalışmaktır. Diyabet rahatsızlığının tahmini için RO, DVM, KA ve LR sınıflandırma algoritmalarını kullanmaktadırlar. Deneysel sonuçlar, RO algoritmasının %84,11 doğruluk oranı ile en iyi tahmin sonucuna sahip olduğunu göstermektedir.

Bu araştırma çalışmasında, Apache Spark'a dayalı dağıtık makine öğrenimi yöntemleri kullanarak diyabet rahatsızlığının tahmini için Java tabanlı bir sistem geliştirilmiştir. Sistemde, tahminlerin yapılabilmesi için farklı makine öğrenimi modelleri oluşturulmuştur. Bu modeller içerisinde YSA,

DVM, LR, RO ve NB olmak üzere beş makine öğrenimi sınıflandırma algoritması kullanılmıştır. Ayrıca, sınıflandırma algoritmaları doğruluk, duyarlılık, özgünlük, kesinlik, negatif tahmin değeri, hata oranı ve f-ölçüsü olmak üzere yedi farklı değerlendirme ölçütü kullanılarak performansları karşılaştırılmıştır.

II. MATERYAL VE METOT

Bu bölümde, Apache Spark büyük veri teknolojisi ile oluşturulan analiz sisteminin yapısı Şekil 1'de gösterilmekte ve süreci alt bölümler halinde açıklanmaktadır.



Şekil 1. *Apache Spark tabanlı oluşturulan sistemin aşamaları*

A. VERİ SETİ

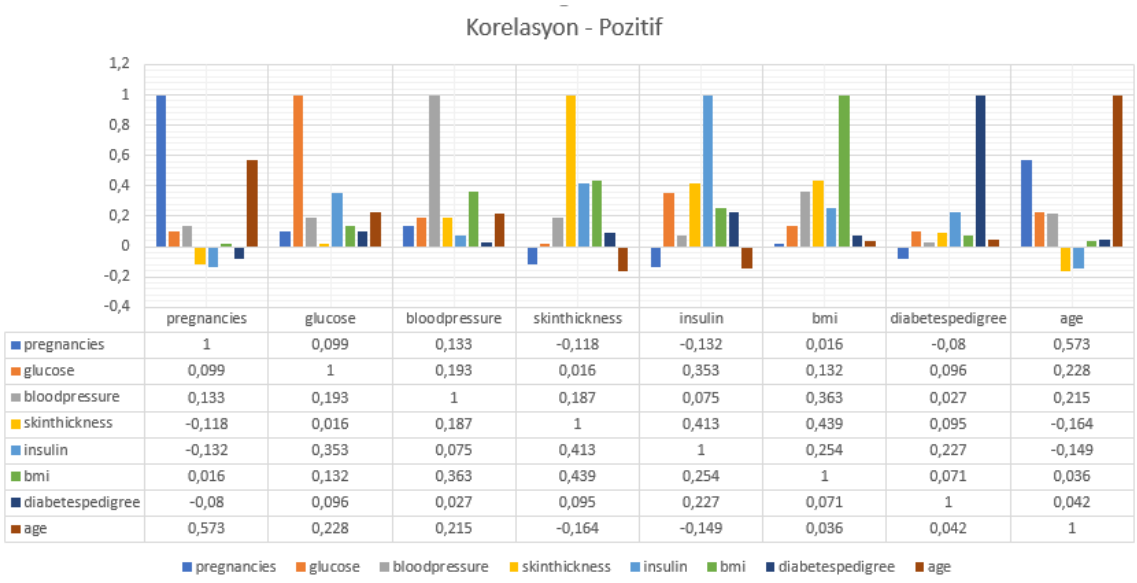
Bu çalışmada, UCI makine öğrenimi laboratuvarında oluşturulan Pima Indians diyabet hastalığı veri seti kullanılmıştır [11]. Bu veri setinde, diyabet rahatsızlığı olan 268 kişi ve olmayan 500 kişinin verileri bulunmakta ve toplamda 768 satırlık örnekten oluşmaktadır. Diyabet rahatsızlığını teşhis etmek için Tablo1'deki gibi 8 farklı parametre dikkate alınmıştır.

Tablo 1. *UCI Pima Indians Veri Seti*

No	Parametre	Açıklama
1	Pregnancies	Kişinin hamilelik sayısı
2	Glucose	Glikoz yoğunluk miktarı
3	Blood Pressure	Kan basıncı
4	Skin Thickness	Deri kıvrım kalınlığı
5	Insulin	İnsülin
6	BMI	Vücut kitle indeksi
7	Diabetes Pedigree	Diyabet soyağacı işlevi
8	Age	Kişinin yaşı

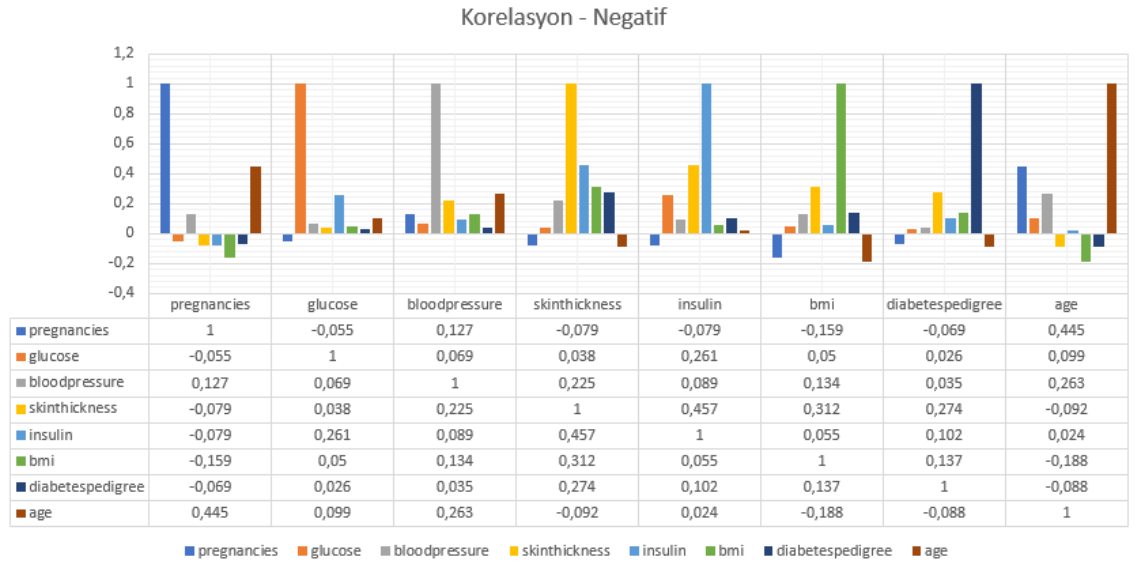
Veri setinde, her bir veri için 8 farklı parametre vardır. Diyabet rahatsızlığı olmayan kişiler pozitif sınıf olarak, diyabet rahatsızlığı olan kişiler ise analiz amacıyla negatif numuneler olarak ele alınmıştır. Sekiz "negatif" ve "pozitif " örneği parametresi arasındaki korelasyon, Şekil 2 ve 3' te gösterildiği gibi bu iki

örnek sınıfının parametreleri arasında pozitif yönde bir korelasyon vardır. Parametreler arasında korelasyonun yüksek olması sınıflandırma algoritmalarının daha iyi performansa sahip olmasını sağlamaktadır.



	pregnancies	glucose	bloodpressure	skinthickness	insulin	bmi	diabetespedigree	age
pregnancies	1	0,099	0,133	-0,118	-0,132	0,016	-0,08	0,573
glucose	0,099	1	0,193	0,016	0,353	0,132	0,096	0,228
bloodpressure	0,133	0,193	1	0,187	0,075	0,363	0,027	0,215
skinthickness	-0,118	0,016	0,187	1	0,413	0,439	0,095	-0,164
insulin	-0,132	0,353	0,075	0,413	1	0,254	0,227	-0,149
bmi	0,016	0,132	0,363	0,439	0,254	1	0,071	0,036
diabetespedigree	-0,08	0,096	0,027	0,095	0,227	0,071	1	0,042
age	0,573	0,228	0,215	-0,164	-0,149	0,036	0,042	1

Şekil 2. *Pozitif sınıflar için doğrusal korelasyon*

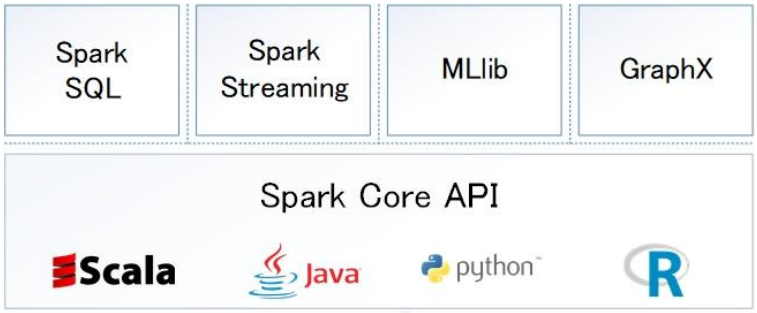


	pregnancies	glucose	bloodpressure	skinthickness	insulin	bmi	diabetespedigree	age
pregnancies	1	-0,055	0,127	-0,079	-0,079	-0,159	-0,069	0,445
glucose	-0,055	1	0,069	0,038	0,261	0,05	0,026	0,099
bloodpressure	0,127	0,069	1	0,225	0,089	0,134	0,035	0,263
skinthickness	-0,079	0,038	0,225	1	0,457	0,312	0,274	-0,092
insulin	-0,079	0,261	0,089	0,457	1	0,055	0,102	0,024
bmi	-0,159	0,05	0,134	0,312	0,055	1	0,137	-0,188
diabetespedigree	-0,069	0,026	0,035	0,274	0,102	0,137	1	-0,088
age	0,445	0,099	0,263	-0,092	0,024	-0,188	-0,088	1

Şekil 3. *Negatif sınıflar için doğrusal korelasyon*

B. APACHE SPARK

Apache Spark, güvenilir, hızlı, yüksek düzeyde hataya dayanıklı ve Hadoop ile tutarlı olacak şekilde oluşturulmuş popüler bir veri işleme çerçevesidir. Spark, çok büyük veri setlerini işleyebilen dağıtık mimari ve analiz için bir çerçeve sunar [12]. Spark'ın en önemli özelliği, bellek içi hesaplama yapabilmesidir. Hesaplama için yavaş olan sabit diske erişmek yerine RAM bellek kullanır. Büyük verileri işlerken işlemleri seri olarak yürütmek yerine iş yükünü paralel olarak dağıtarak çalışır [13]. Ayrıca Spark, SQL için Spark SQL, makine öğrenimi için MLlib, grafik işleme için GraphX ve Spark Streaming dahil olmak üzere zengin üst düzey kütüphaneleri destekler. Spark'ın ekosistemi Şekil 4'te gösterilmektedir.



Şekil 4. *Apache Spark ekosistemi*

Apache Spark, Hadoop öğeleriyle uyumludur. Ayrıca, YARN'nı kullanarak bir Hadoop kümesinde çalışabilir. Apache Spark, Scala dilinde yazılmıştır. Ancak Scala, Java, SQL, Python ve R gibi dilleri de desteklemektedir. Bununla birlikte, MLlib, diğer ML kütüphanesi ile içerisinde bulunan farklı makine öğrenimi algoritmaları kullanılarak akıllı şehirler [14], afet yönetimi [15], finans [16] ve sağlık [17] gibi alanlara yönelik analizler yapılabilmektedir [18]. Bu çalışmada da Java programlama dili kullanılarak ML kütüphanesindeki makine öğrenimi sınıflandırma algoritmaları ile diyabet rahatsızlığının tahmini yapılmaktadır.

C. MAKİNE ÖĞRENMESİ SINIFLANDIRMA ALGORİTMALARI

Diyabet rahatsızlığını tahmin etmek için makine öğrenimi sınıflandırma modelleri oluşturulmuştur. Veri setini eğitmek ve test etmek için yapay sinir ağları, destek vektör makinaları, lojistik regresyon, rastgele orman ve naive bayes olmak üzere beş makine öğrenimi algoritması uygulanmıştır. Ayrıca, makine öğrenimi modellerini uygulamak için Apache Spark'a ait MLlib kütüphanesi kullanılmıştır. Çalışmamızda, eğitim verileri kullanılarak sınıflandırma modelleri eğitilmektedir. Modelin değerlendirmesi için test verileri kullanılmaktadır.

C. 1. Yapay Sinir Ağları

YSA, bir soruna çözüm üretmek için çok sayıda basit, yüksek oranda birbirine bağlı işlem elemanlarından (nöronlar) oluşan matematiksel bir modeldir. Bir sinir ağının bir avantajı, yeterli sayıda gizli düğüm verildiğinde çeşitli yanıt yüzeylerini modelleyebilmesidir [19]. Bu çalışmada çok katmanlı algılayıcı ağ modeli kullanılmıştır. Ayrıca, bir girdi katmanı, iki gizli katman ve bir çıktı katmanı yer alır.

C. 2. Destek Vektör Makinesi

DVM, Boser, Guyon ve Vapnik tarafından 1992 yılında ortaya çıkan bir sınıflandırma yöntemidir [20]. DVM, büyük hacimli veriler üzerinde yüksek doğruluk elde etmesinden dolayı yaygın olarak kullanılmaktadır. DVM algoritması, veri kümesini eğitim örnekleriyle tutarlı bir şekilde önceden tanımlanmış farklı sayıdaki sınıfa ayıran bir hiper düzlem bulmayı amaçlamaktadır [21]. Çalışmamızda DVM, diyabet rahatsızlığının olması ve olmaması olarak iki sınıflı bir probleme odaklanmaktadır ve doğrusal destek vektör makinesi modeli kullanılmaktadır.

C. 3. Lojistik Regresyon

LR, makine öğrenimi uygulamalarında yaygın olarak kullanılmaktadır. Veri seti verildiğinde, bir olayın meydana gelip gelmeyeceğini tahmin etmeye çalışan en uygun sınıflandırma algoritmalarından biridir. Normalde, yalnızca iki sonuç olduğunda kullanılır: ya olay olur ya da olmaz. Bir dizi girdi verildiğinde, bir olayın meydana gelme olasılığını hesaplayacak bir model oluşturmaya çalışır [22]. Çalışmamızda, model bir değişken vektörünü kavrar ve her girdi değişkeni için katsayıları veya ağırlıkları değerlendirir ve ardından belirtilen kişinin diyabet rahatsızlığının olup olmadığını tahmin eder.

C. 4. Rastgele Orman

RO, denetimli makine öğrenmesi algoritması yöntemlerinden biridir. Denetimli bir öğrenme algoritması, etiketli eğitim verilerinden öğrenir, öngörülemeyen veriler için sonuçları tahmin etmenize yardımcı olur. Algoritmada kullanılan ağaç sayısı ile doğruluk arasında pozitif bir ilişki vardır. Ağaç sayısı arttıkça algoritmanın doğruluk artmaktadır. RO algoritması ile KA algoritması arasındaki fark, RO'da kök düğümü bulmanın ve düğümleri bölmenin rastgele olması ve karar ağacında süreç için olasılıklı bir hesaplamanın var olmasıdır [23]. Çalışmamızın model oluşturma parametrelerinde ağaç sayısı 20, maksimum ağaç derinliği 10, özellik ayrıştırma bölme sayısı 10 ve Gini indeksi kullanılmaktadır.

C. 5. Naive Bayes

NB algoritması, veri seti üzerinde yaptığı tahminlerde Bayes kuralını kullanan sezgisel bir yöntemdir [24]. NB, ölçeklendirilebilen bir sınıflandırma algoritmasıdır. Bir öğrenme sürecinde, doğrusal değişkenleri olan çok sayıda parametre gerektirir. Eğitim verilerinin tek bir yinelemesinde, algoritma verilerinde etiketin her bir özelliğinin koşullu olasılık dağılımını bulmak için Bayes teoremi uygular ve tahmin için kullanır.

D. PERFORMANS DEĞERLENDİRMESİ

Diyabet rahatsızlığı olan veya olmayan kişilerin sağlık değerlerine ait örneklerin sınıflandırılması için beş farklı denetimli makine öğrenimi sınıflandırma algoritması uygulanmıştır. Değerlendirme yapılırken 10-kat çapraz doğrulama, sınıflandırma algoritmalarının performansları için kullanılmıştır. Çalışmada kullanılan sınıflandırma modelleri, doğruluk, duyarlılık, özgünlük, kesinlik, negatif tahmin değeri (NTD), hata oranı, f ölçüsü olmak üzere yedi farklı ölçüm ile değerlendirilmiştir.

Diyabet rahatsızlığı olmayan kişiler pozitif sınıf, diyabet rahatsızlığı olan kişiler ise negatif sınıf olarak kabul edilmiştir. Burada,

Gerçek pozitif (TP) - diyabet rahatsızlığının olmadığı tahmin edilen, gerçekte de diyabet rahatsızlığı olmayan kişilerin sayısı.
Yanlış pozitif (FP) - diyabet rahatsızlığı olmadığı tahmin edilen, ancak gerçekte diyabet rahatsızlığı olan kişilerin sayısı.
Gerçek negatif (TN) - diyabet rahatsızlığının var olduğu tahmin edilen, gerçekte de diyabet rahatsızlığı olan kişilerin sayısı.
Yanlış negatif (FN) - diyabet rahatsızlığının var olduğu tahmin edilen, ancak gerçekte diyabet rahatsızlığı olmayan kişilerin sayısı.

Sınıflandırma doğruluğu: Veri setindeki hastalığın doğru bir şekilde tahmin edilme olasılığıdır. Doğru şekilde tahmin edilen örneklerin toplam örnek sayısına oranıdır.

$$\text{Doğruluk} = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Duyarlılık veya geri çağırma, tahmin edilen pozitif örneklerin toplam pozitif örneklere oranıdır.

$$\text{Duyarlılık} = \frac{TP}{TP + FN} \quad (2)$$

Özgünlük, tahmin edilen negatif örneklerin toplam negatif örneklere oranıdır.

$$\text{Özgünlük} = \frac{TN}{TN + FP} \quad (3)$$

Kesinlik, pozitif olarak sınıflandırılan tüm örneklerle, gerçek pozitifin oranıdır.

$$\text{Kesinlik} = \frac{TP}{TP + FP} \tag{4}$$

NTD, doğru tahmin edilen negatifin (diyabet rahatsızlığının varlığı) toplam tahmin edilen negatife oranı (diyabet rahatsızlığının varlığı olarak sınıflandırılan toplam örnek).

$$\text{NTD} = \frac{TN}{TN + FN} \tag{5}$$

Hata oranı, yanlış sınıflandırılmış örneklerin toplam örnek sayısına oranı, hata oranı olarak bilinir. Diyabet rahatsızlığının olması durumu, diyabet rahatsızlığı olmaması olarak sınıflandırılırsa (E1), tip I hata olarak tanımlanır. Diyabet rahatsızlığının olmaması, diyabet rahatsızlığının olması olarak sınıflandırılırsa (E2), tip II hata olarak sınırlandırılır.

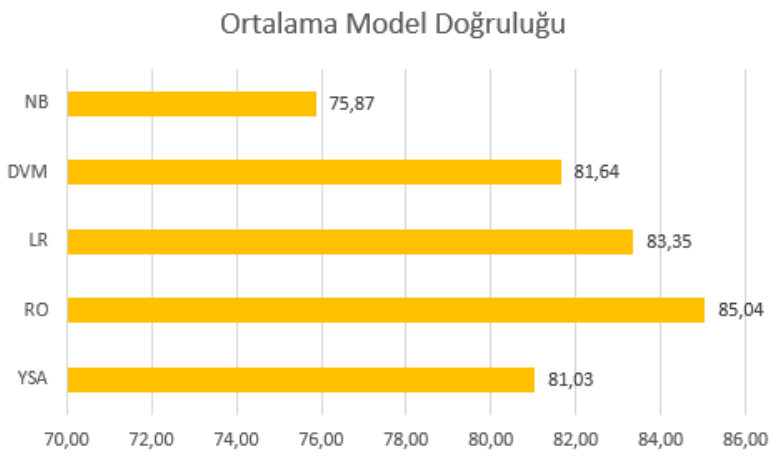
$$\text{Hata oranı} = \frac{E1 + E2}{Toplam\ örnek\ sayısı} \tag{6}$$

F ölçüsü, kesinlik ve duyarlılık arasındaki harmonik ortalamadır. En iyi performans için 1, en kötü performans için ise 0 değeri verilir.

$$F_1 = \frac{2 * Kesinlik * Duyarlılık}{Kesinlik + Duyarlılık} \tag{7}$$

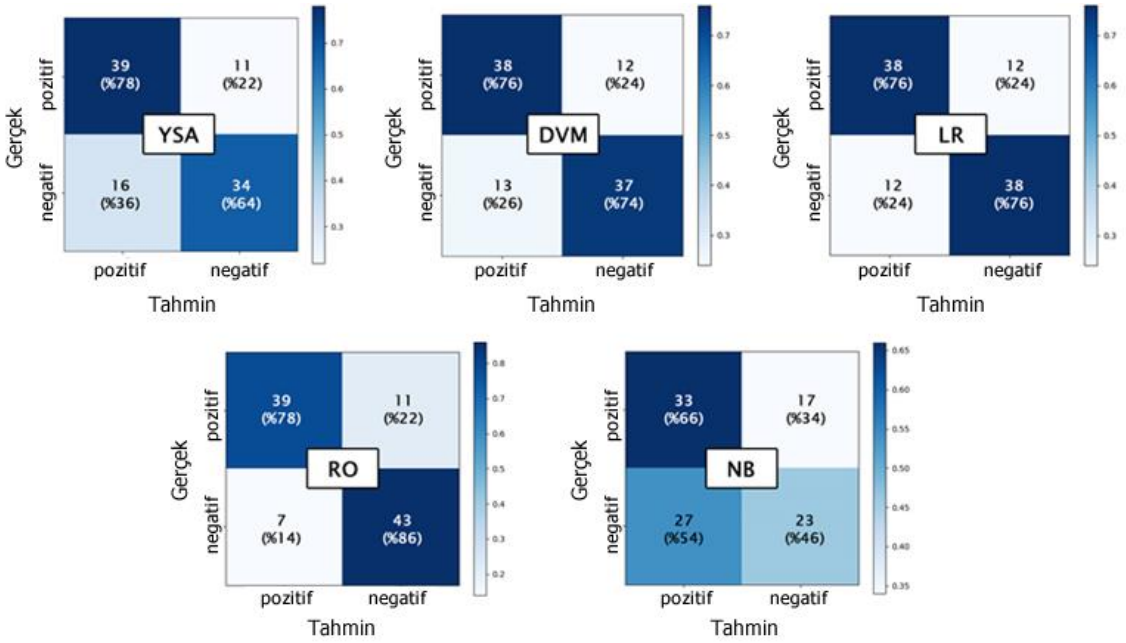
III. BULGULAR

Çalışmada, Apache Spark büyük veri işleme teknolojisi kullanılarak diyabet rahatsızlığı tahmini yapılmaktadır. Rahatsızlık tahmini için Apache Spark içerisinde bulunan MLlib kütüphanesine ait beş farklı makine öğrenimi sınıflandırma algoritması kullanılmaktadır. Diyabet rahatsızlığı için gerçekleştirilen uygulama, Java programlama dili kullanılarak geliştirilmiştir. Bununla birlikte, sınıflandırma algoritmalarının performansları materyal ve yöntem bölümünde belirtilen yedi farklı ölçüte göre değerlendirilmiştir. Kullanılan veri setinde 268'si diyabet rahatsızlığı bulunan ve 500'ünde diyabet rahatsızlığı bulunmayan toplam 768 örnek bulunmaktadır. Ancak, modelin tahmininde sapmaların yaşanmaması için veri setindeki sınıflara ait örnekler ön işleme aşamasında eşitlenmiştir. Bu nedenle, çalışmamızda 268 pozitif ve 268 negatif sınıfa ait örnekler kullanılmıştır. Aynı zamanda, modelin oluşturulması için veri setindeki 436 örnek eğitim verisi olarak, 100 örnek ise test verisi olarak kullanılmıştır. Ek olarak, modelin doğrulanmasında 10-kat çapraz doğrulama kullanılmıştır. Veriler rastgele on eşit parçaya bölünür ve her adımda bir tanesi test verisi, geri kalan parçalar ise eğitim verisi olarak kullanılır.



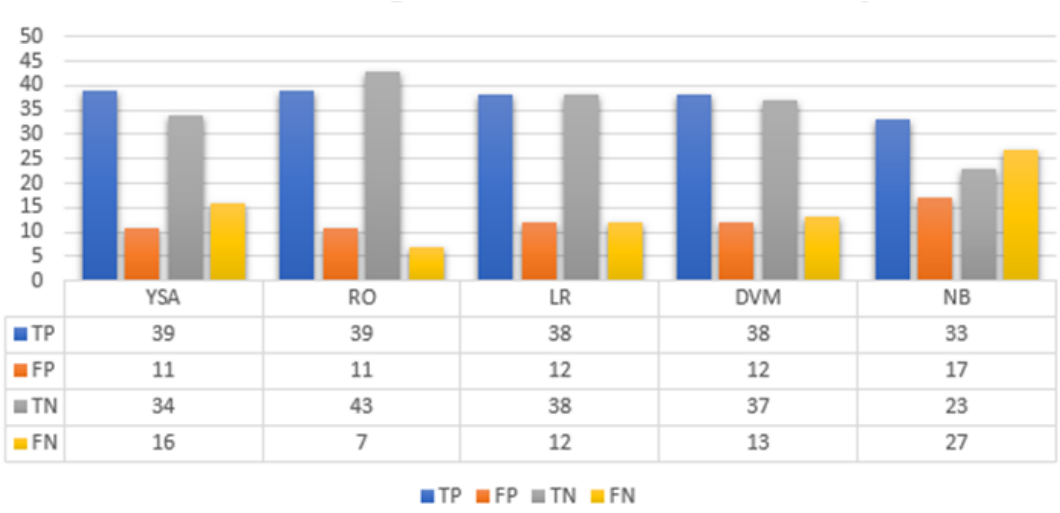
Şekil 5. *Oluşturulan modellerin ortalama doğruluk oranları*

Eğitim verileri ile oluşturulan modelde, 10-kat çapraz doğrulama yöntemi kullanılarak modelin doğruluğu incelenmiştir. Şekil 5 incelendiğinde, elde edilen sonuçlar modellerin ortalama doğruluk oranlarıdır. Bu bağlamda, en iyi doğruluk oranı RO algoritması ile oluşturulan modele aittir. Ayrıca, NB algoritması ile oluşturulan modelin en düşük doğruluk oranına sahip olduğu görülmektedir.



Şekil 6. *Sınıflandırma algoritmalarının karmaşıklık matrisleri*

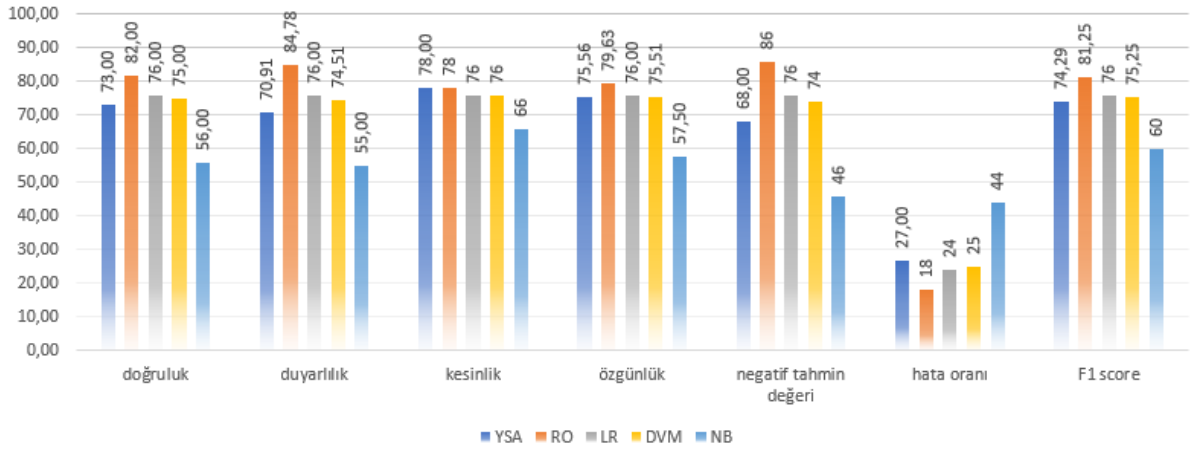
YSA, DVM, LR, RO ve NB algoritmaları için tahmin sonuçlarının karmaşıklık matrisleri Şekil 6'da gösterilmektedir. Şekil 7'de ise bu makine öğrenimi sınıflandırma algoritmalarının tahminlerinin sonuçları gösterilmektedir. Sonuçlar incelendiğinde, YSA ve RO'ın en yüksek sayıda gerçek pozitifi (diyabet rahatsızlığının olmadığı), RO'ın ise en yüksek sayıda gerçek negatifi (diyabet rahatsızlığının olduğu) öngördüğü açık bir şekilde görülmektedir. Bununla birlikte NB'in ise en düşük gerçek pozitif ve negatifi öngördüğü görülmektedir.



	YSA	RO	LR	DVM	NB
TP	39	39	38	38	33
FP	11	11	12	12	17
TN	34	43	38	37	23
FN	16	7	12	13	27

Şekil 7. *Makine öğrenimi sınıflandırma algoritmalarının tahmin sonuçları*

Şekil 8, RO'ın %82'lik maksimum sınıflandırma doğruluğu ile diğer tüm makine öğrenimi yöntemlerinden daha iyi performans gösterdiğini belirtirken, ikinci en yüksek sınıflandırma doğruluğunun %76 ile LR'un elde ettiğini belirtmektedir. Bununla birlikte, RO'ın %84,78, %78, %79,63 ve %81,25'lik en yüksek duyarlılığa, kesinlik, özgünlük değerine ve F1 skoruna sahip olduğu gösterilmektedir. Ek olarak, RO'ın %86'lık negatif tahmin değeri dikkat çekmektedir. Bu değer ile RO, diyabet rahatsızlığı olan kişilerin tahmininde diğer algoritmalara göre yüksek bir performans göstermiştir. Diyabet rahatsızlığının tahmininde kullanılan ve en düşük doğruluk (%60), duyarlılık (%70,5), kesinlik (%63), özgünlük (%46) ve F1 skoru (%66,5) değerleri ile NB sınıflandırıcısı en kötü performansa sahiptir.



Şekil 8. *Makine öğrenimi sınıflandırma algoritmalarının çeşitli değerler bakımından performans değerleri*

IV. SONUÇ

Diyabet rahatsızlığının tahmini, insanların hayatını kurtarabilmekte ve tedavisi üzerinde önemli bir etkiye sahip olabilmektedir. Bu çalışma, diyabet rahatsızlığının olup olmadığını tahmin etmek için Apache Spark büyük veri işleme teknolojisine ve ona ait makine öğrenimi yöntemlerine dayanan bir iş akışı sağlamaktadır. Bu bağlamda, diyabet rahatsızlığının tahmini için birçok parametreye dayanan beş farklı sınıflandırma algoritması kullanılmaktadır. Bu algoritmalara dayalı modeller, on kat çapraz

doğrulama kullanılarak doğrulanmış ve doğruluk, duyarlılık, kesinlik, özgünlük, negatif tahmin değeri, hata oranı ve f-ölçüsü performans ölçümleri açısından değerlendirilmiştir.

Sonuçlar incelendiğinde, RO sınıflandırma algoritmasının diğer algoritmalara göre daha iyi performans verdiği saptanmıştır. NB algoritması ise kullanılan diğer algoritmalara göre çok daha düşük performansa sahiptir. Bunun nedeni, veri setindeki parametreler arasında oluşan pozitif korelasyon olduğu söylenebilir. Ancak, NB de, tahmin aşamasında, parametreler bağımsız olarak değerlendirilmektedir. Diğer algoritmalarda ise parametreler arasında ne kadar yüksek korelasyon var ise o kadar iyi sonuçlar elde edilmektedir. Bu nedenle, çalışmalarda korelasyonu yüksek bir veri seti kullanıldığında tahmin sonuçları çok daha iyi olabilmektedir.

Bu çalışma, daha önce yapılan araştırmalar ile karşılaştırıldığında, Han vd. (2008) çalışmasındaki NB'in performansı (%76,30) ve Barakat vd. (2010) çalışmasında DVM'nin performansı (%94), çalışmamızda kullanılan NB (%56) ve DVM (%75) algoritmalarına göre daha iyi performans göstermektedirler. Ancak çalışmamızda en yüksek performansa sahip olan RO (%82), Zou vd (2018) çalışmasında belirtilen RO'ın performansından (%80) daha iyi olduğu görülmektedir.

Sonuç olarak, Apache Spark ve içerisinde kullanılan yetenekli kütüphaneleri ile diyabet gibi diğer rahatsızlıkların tahmininde kullanılabilme potansiyeline sahiptir. Gelecekte bu çalışma, daha da genişletilerek Apache Spark'ın gerçek zamanlı veri işleme özelliği ile insan vücuduna bağlı çeşitli sensörler kullanılarak alınan daha yüksek hacimli veriler analiz edilerek diyabet rahatsızlığının erken teşhisinin yapılması planlanmaktadır.

V. KAYNAKLAR

[1] World Health Organization. (2021, June 15). *WHO Diabetes Program* [Online]. Erişim: https://www.who.int/health-topics/diabetes

[2] Apache Flink. (2021, June 15). *Apache Flink* [Online]. Erişim: https://flink.apache.org/

[3] Apache Hadoop. (2021, June 15). *Apache Hadoop* [Online]. Erişim: https://hadoop.apache.org/

[4] Apache Spark. (2021, June 15). *Apache Spark* [Online]. Erişim: https://spark.apache.org/

[5] J. Han, J.C. Rodriguez, J.C., and M. Beheshti, "Discovering decision tree based diabetes prediction model," in *Advances in Software Engineering*, 1st ed., Hainan Island, China: Springer, 2008, pp. 99-109.

[6] P.S. Kumar, and S. Pranavi, "Performance analysis of machine learning algorithms on diabetes dataset using big data analytics," *International Conference on Infocom Technologies and Unmanned Systems*, Dubai, UAE, 2017, pp. 508-513.

[7] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers in Genetics*, vol. 9, no. 515, pp. 1-10, 2018.

[8] N.H. Barakat, A.P. Bradley, and M.N. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 4, pp. 1114-1120, 2010.

[9] A. Mir, and S.N. Dhage, "Diabetes disease prediction using machine learning on big data of healthcare," 4*th International Conference on Computing Communication Control and Automation*, Pune, India, 2018, pp. 1-6.

[10] F. Hassan and M.E. Shaheen, "Predicting diabetes from health-based streaming data using social media, machine learning and stream processing technologies," *International Journal of Engineering Research and Technology*, vol. 13, no. 8, pp. 1957-1967, 2020.

[11] Kaggle. (2021, June 15). *Pima Indians Diabetes Database* [Online], Erişim: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[12] M. Zaharia, M. Chowdhury, T. Das, A Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica, "Resilient distributed datasets: a fault-tolerant abstraction for in-memory cluster computing," *9th Symposium on Networked Systems Design and Implementation*, California, USA, 2012, pp. 15-28.

[13] X. Meng, J. Bradley, B. Yavuz, E. Sparks, S. Venkataraman, D. Liu, J. Freeman, D.B. Tsai, M. Amde, S. Owen and D. Xin, "MLlib: machine learning in apache spark," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 1235-1241, 2016.

[14] S. Ameer, M.A. Shah, A. Khan, H. Song, C. Maple, S. Islam, and M.N. Asghar. "Comparative analysis of machine learning techniques for predicting air quality in smart cities," *IEEE Access*, vol. 7, no. 2019, pp. 128325-128338, 2019.

[15] K. Kucuk, C. Bayilmis, A.F. Sonmez, and S. Kacar. "Crowd sensing aware disaster framework design with IoT Technologies," *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 4, pp. 1709-1725, 2020.

[16] X. Tian, R. Han, L. Wang, G. Lu, and J. Zhan. "Latency critical big data computing in finance," *The Journal of Finance and Data Science*, vol. 1, no. 1, pp. 33-41, 2015.

[17] L.R. Nair, S.D. Shetty, and S.D. Shetty. "Applying spark based machine learning model on streaming big data for health status prediction," *Computers & Electrical Engineering*, vol. 65, no. 393-399, 2018.

[18] M. Alber, "Masterarbeit: big data and machine learning: a case study with bump boost", Department of Smart Systems and Robotics, Master Thesis, Freie University, Berlin. Germany, 2014.

[19] J.K. Basu, D. Bhattacharyya and T.H. Kim, "Use of artificial neural network in pattern recognition," *International Journal of Software Engineering and Its Applications*, vol. 4, no. 2, pp. 23-34, 2010.

[20] B. E. Boser, I. M. Guyon, and V.N. Vapnik, "A training algorithm for optimal margin classifiers," *5th Annual ACM Workshop on Computational Learning Theory*, Pittsburgh, ABD, 1992, pp. 144-152.

[21] G. Zhu, and D. G. "Blumberg. classification using aster data and svm algorithms; the case study of beer sheva, israel," *Remote Sensing of Environment*, vol. 80, no. 2, pp. 233-240, 2002.

[22] D.W. Hosmer Jr, S. Lemeshow and R.X. Sturdivant, "Introduction to the logistic regression model", *Applied Logistic Regression*, 3rd ed., New Jersey, USA: John Wiley & Sons, 2013, vol. 398, pp. 1-35.

[23] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5-32, 2001.

[24] P. Langley, W. Iba, and K. Thompson, "An analysis of bayesian classifiers," *Proceedings of The Tenth National Conference on Articial Intelligence*, California, USA, 1992, pp. 223-228.