Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 32, S. 7-13, Aralık 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 32, pp. 7-13, December 2021

Research Article

Destek Vektör Makineleri ve Naive Bayes Sınıflandırma Algoritmalarını Kullanarak Diabetes Mellitus Tahmini

Güneş Harman[*],

[1*] Yalova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Yalova, Türkiye, (ORCID: 0000-0001-5413-124X), gunes.guclu@yalova.edu.tr

(International Conference on Design, Research and Development (RDCONF) 2021 – 15-18 December 2021)

(**DOI:** 10.31590/ejosat.1041186)

ATIF/REFERENCE: Harman, G. (2021). Destek Vektör Makineleri ve Naive Bayes Sınıflandırma Algoritmalarını Kullanarak Diabetes Mellitus Tahmini. *Avrupa Bilim ve Teknoloji Dergisi*, (32), 7-13.

Öz

Makine öğrenmesi, herhangi bir insan müdahalesi olmadan elde olan verilerden veya analizlerinden daha iyi sonuçlar elde edilmesine yardımcı olan alanlardan biridir. Ciddi ve karmaşık durumları analiz etmek ve doğruluk oranı yüksek tahminlerde bulunmak için son yıllarda gelişen teknolojiyle birlikte özellikte tıbbi teşhis alanında yaygın olarak kullanılmaktadır. Bu çalışma kapsamında Pima Indians Diyabet veri seti (Pima Indian Diabetes Dataset) üzerinde Naive Bayes ve Destek Vektör Makineleri (DVM) makine öğrenme algoritmaları kullanılarak diyabet hastalığı erken evrede teşhis edilmeye çalışılmıştır. Kullanılan sınıflandırıcıların performanslarını artırmak için veri setinde eksik değerler çarpıklık durumuna göre tekrar yapılandırılmış, veri standardizasyonu standart ölçeklendirme kullanılarak yapılmıştır. Ayrıca sınıf dengesizlik probleminin sınıflandırma üzerindeki olumsuz etkisini azaltmak için Sentetik Azınlık Aşırı-Örnekleme (SMOTE) tekniği kullanılmıştır. Çalışma kapsamında oluşturulan sınıflandırıcıların değerlendirme kriterleri Doğruluk Oranı (Accuracy Rate), Kesinlik (Precision), Duyarlılık (Recall) ve F1-Skore (F1 Score) metrikleri kullanılarak hesaplanmıştır. Destek Vektör Makineleri %90 doğruluk oranı ile en iyi sunucu vermiştir.

Anahtar Kelimeler: Makine öğrenmesi, Sınıflandırma, Destek vektör makineleri, Naive bayes, Diyabet.

Prediction of Diabates Mentilus by Using SVM and Naive Bayes Classification Algorithms

Abstract

Machine learning is one of the fields that help to get better results from data or analysis without any human intervention. In recent years with the developing technology, it is widely used in the field of medical diagnosis, especially to analyze serious and complex situations and make predictions with high accuracy. In this study, it was tried to diagnose diabetic disease at an early stage by using Naïve Bayes and Support Vector Machines (DVM) machine learning algorithms on Pima Indians Diabetes Dataset. In order to increase the performance of the classifiers used, the missing values in the data set were restructured according to the skewness, and data standardization was done using standard scaling. Then, Synthetic Minority Oversampling (SMOTE) technique was used to reduce the negative effect of class imbalance problem on classification. Evaluation criteria of the classifiers created within the scope of the study were calculated by using Accuracy Rate, Precision, Recall and F1-Score (F1 Score) values. According to these results, Support Vector Machines gave the best server with 88% accuracy rate.

Keywords: Machine learning, Classification, Support vector machine, Naïve bayes, Diabetes.

[*] Güneş HARMAN: gunes.guclu@yalova.edu.tr

1. Giriş

Şeker hastalığı yani tıp dilinde Diabetes Mellitus olarak adlandırılan diyabet hastalığı, kan şekerinin yükselmesine neden olan en ölümcül ve kronik hastalıklardan biri olarak kabul edilmektedir. Kan şekerini düzenleyen insülin hormonunun eksikliği, yeterince kullanılamaması veya üretilememesi durumlarında ortaya çıkmaktadır. Dünya Sağlık Örgütü'nün (DSÖ-World Health Organization) son yıllarda yayınlamış olduğu göstergelere bakıldığında dünya genelinde yaklaşık olarak 422 milyon insanın diyabet hastası olduğu ve her yıl meydana gelen ölümlerin yaklaşık olarak 1.6 milyonun doğrudan diyabete bağlı olduğu açıklanmıştır [1]. Aynı zamanda 2015 ve sonrası Uluslararası Diyabet Federasyonu (IDF) verilerine bakıldığında dünyada 415 milyon birey diyabet hastası iken bu sayının 2040 yılında %55 artarak 642 milyona ulaşacağı tahmin edilmektedir. Bu durum diyabet hastalığının tüm dünya popülasyonuna ve tüm yaş gruplarına yayılmış, yaygın ve hızlı artan bir çeşit hastalık olduğunu göstermektedir.

Diyabet hormonal duruma bağlı olmasından dolayı ömür boyu süren hastalıklardan biridir. Şeker hastalığı, başta böbrek fonksiyonları ve tansiyon olmak üzere vücut genelinde ciddi tahribata yol açmaktadır. Hastalığın erken teşhis edilmesi ve tedaviye başlanması yani zamanında tedbir alınması beraberinde ve sonrasında oluşacak diğer hastalıkların önüne geçmek ve engellemek için çok büyük önem arz etmektedir. Son yıllarda teknolojinin gelişmesiyle birlikte özellikle tıbbi teşhis alanında makine öğrenmesi yöntemleri kullanılmaktadır. Makine öğrenimi, herhangi bir insan müdahalesi olmadan verilerden ve analizlerinden daha iyi öğrenmeye yardımcı olan, yaygın olarak büyüyen bir alandır. Ciddi ve karmaşık durumları analiz etmek ve tespit etmek için özellikle sağlık hizmetleri alanında popüler bir şekilde kullanılmaktadır. Makine öğrenmesinde kullanılan sınıflandırma algoritmaları yüksek oranda doğruluk sonuçları vermektedir. Bu durum daha hızlı karar verme ve hekimlere yardımcı olma açısından çok önemlidir. Diğer bütün hastalıklarda olduğu gibi şeker hastalığının erken teşhis ve tedavi süreci hayat kurtarmakla birlikte kişilerin yaşam kalitesini daha iyi hale getirmektedir.

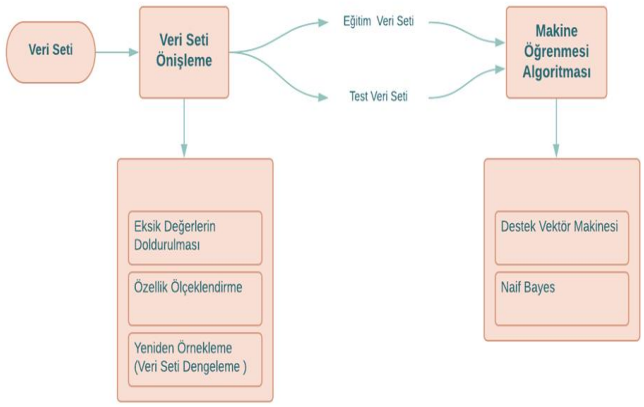
Literatüre baktığımızda Diyabet hastalığın makine öğrenmesi yöntemleri kullanılarak sınıflandırılması için çok farklı algoritma ve yöntemler kullanılmıştır. Bunun yanı sıra farklı veri setleri de kullanılmıştır. En yaygın kullanılanlardan biri de "Pima Indians Diabetes" veri seti kümesidir. Bu veri seti kullanılarak yapılan çalışmalardan biri olan [2] şeker hastalığın tahmin edilmesi için Naive Bayes, Rastgele Orman, Karar Ağaçları, Lojistik Regresyon ve k-En yakın komşuluk makine öğrenmesi sınıflandırma algoritmaları kullanılmıştır. Sınıflandırma algoritmaları karşılaştırıldığında %89 doğruluk oranıyla en iyi, sonucu Rastgele Orman Algoritması vermiştir. Diğer bir çalışmada [3] yine aynı veri seti üzerinde Karar Ağaçları, Naive Bayes ve Destek Vektör Makineleri olmak üzere üç farklı makine öğrenmesi yöntemi kullanılarak diyabet hastalığı erken evrede teşhis edilmeye çalışılmıştır. Kullanılan algoritmaların performans değerlendirmesi Hassasiyet, Doğruluk, F-Skore ve Eğri Altında Kalan Alan metrikleri kullanarak yapılmıştır. Elde edilen sonuçlara göre %76,30 doğruluk oranıyla en iyi performansı Naive Bayes algoritması vermiştir. Aynı veri seti kullanılarak yapılan çalışmalardan bir diğeri [4] Çok Katmanlı Yapay Sinir Ağları, Radyal Temel Fonksiyonu ve Genel Regresyon Sinir Ağı olmak üzere diyabet hastalığının teşhis edilmesinde üç farklı sinir ağ yapısı kullanmıştır. En iyi sonucu %81 oranıyla Genel Regresyon Sinir Ağları vermiştir. Bir diğer çalışmada [5] Bangladeş Sylhet Diyabet Hastanesin'den elde edilen veriler üzerinde 6 farklı makine öğrenmesi yöntemi kullanılmıştır. Çalışma kapsamında kullanılan Çok Katmanlı Algılayıcı Yapay Sinir Ağları, Destek Vektör Makinaları, Karar Ağaçları, Topluluk Öğrenme Algoritmaları, Doğrusal Ayrımcı Analizi ve k-NN metotları arasında en iyi sonucu %99,81 doğruluk oranı ile k-NN algoritması vermiştir. Bir diğer çalışmada [6] Ulusal Sağlık ve Beslenme İnceleme Anketinden (National Health and Nutrition Examination Survey) elde edilen 2009–2012 yıllarında yürütülen diyabet veri setini kullanılmıştır. Lojistik regresyon (LR), diyabet hastalığı için risk faktörlerini *p* değeri ve olasılık oranına (OR) dayalı olarak belirlemek için kullanılmıştır. Diyabetik hastaları tahmin etmek için Naive Bayes, Karar Ağacı, Adaboost ve Rastgele Orman algoritmaları sınıflandırma modeli olarak kullanılmıştır. Kullanılan sınıflandırma algoritmalarının performansları, doğruluk oranı ve eğri altındaki alan kullanılarak değerlendirilmiş ve en iyi sonucu %94.25 doğruluk oranı ile Rastgele Orman algoritması vermiştir.

Bu çalışmanın temel amacı, diyabet hastalığının teşhis edilmesi için farklı makine öğrenmesi sınıflandırma algoritmaları yaklaşımları kullanılmasıdır. Çalışmada kullanılan veri setim Ulusal Diyabet ve Sindirim ve Böbrek Enstitüsü'nden (National İnstitute of Diabetes and Digestive and Kidney Diseases) alınan, 21 yaş ve üstü kadınlar için olan Pima Indians Diyabet veri seti kümesidir. Kullanılan veri seti içerisinde diyabet hastalığının teşhisinde yer alan belirli tanısal ölçümlere dayalı veriler bulunmaktadır. 768 kayıtlı ve 9 özniteliğe sahip olan veri seti üzerinde Gaussian Naïve Bayes ve Destek Vektör Makineleri sınıflandırma algoritmaları kullanılarak diyabet hastalığının teşhisinde en iyi ve doğru sonucun elde edilmesi amaçlanmıştır. Kullanılan sınıflandırıcıların performanslarını artırmak için veri setinde içerisindeki eksik değerler çarpıklık durumuna göre tekrar yapılandırılmış, veri standardizasyon standart ölçeklendirme kullanılarak yapılmıştır. Aynı zamanda sınıf dengesizlik probleminin sınıflandırma üzerindeki olumsuz etkisini azaltmak için Sentetik Azınlık Aşırı-Örnekleme (SMOTE) tekniği kullanılmıştır. Çalışma kapsamında oluşturulan sınıflandırıcıların değerlendirme kriterleri Doğruluk Oranı (Accuracy Rate), Kesinlik (Precision), Duyarlılık (Recall) ve F1-Skore (F1 Score) değerleri kullanılarak hesaplanmıştır. Bu sonuçlara göre %88 doğruluk oranı ile en iyi sunucu Destek Vektör Makineleri vermiştir.

Çalışmanın ilerleyen bölümleri şu şekilde düzenlenmiştir. Bölüm 2, çalışma süresince kullanılan veriler üzerinde uygulanan yöntemler ve analizler hakkında ayrıntılı bilgi verilmiştir. Bölüm 3'te, kullanılan makine öğrenmesi algoritmaları hakkında kısa açıklamalar bulunmaktadır. Bölüm 4, Python kodlama dili kullanarak hazırlanan sınıflandırma algoritmalarının uygulanması ve performans değerlendirme ölçütlerine yer verilmiştir. Çalışmanın son kısmı olan Bölüm 5'te açıklamalar ve sonuçlar bulunmaktadır.

2. Materyal ve Metot

Bu bölüm yapılmış olan çalışmayı gerçekleştirmek için kullanılan yaklaşımları açıklayan metodolojiyi içerir. Çalışma için kullanılan model diyagramı ana hatlarıyla Şekil 1'de gösterilmiştir.



Şekil 1. Kullanılan model diyaramı.

2.1. Diyabet Veri Seti

Çalışma kapsamında kullanılan veri seti Ulusal Diyabet ve Sindirim ve Böbrek Enstitüsü'nden (National İnstitute of Diabetes and Digestive and Kidney Diseases) alınan, 21 yaş ve üstü kadınlar için olan Pima Indians diyabet veri setidir. Kullanılmış olan veri setinin amacı, bir hastanın diyabetli olup olmadığını, veri setine dâhil edilen belirli tanısal ölçütlere dayanarak tahmin etmektir. Kullanılan veri seti 268'i diyabet hastası, 500'ü diyabet hastası olmayan toplam 768 kayıttan ve 9 nitelikten oluşmaktadır. (8 öznitelik ve 1 sınıf değişkeni). Veri setinde bulunan niteliklere ait bilgiler Tablo 1'de ayrıntılı olarak verilmiş.

Tablo 1.Veri setinde bulunan özellikler.

Sayı	Nitelik	Açıklama
X1	Gebelik (Pregnancies)	Hamile kalma sayısı
X2	Glikoz (Glucose)	Plazma glukoz konsantrasyonu (2 saat oral glukoz tolerans testi)
X3	Kan Basıncı (Tansiyon) (Blood Pressure)	Kan Basıncı (mm/Hg)
X4	Cilt Kalınlığı (Skin Thickness)	Deri Kıvrım Kalınlığı (mm)
X5	İnsülin (Insulin)	2 saatlik insülin serum (mu U/ml)
X6	Vücut kitle indeksi (BMI)	Vücut Kitle İndeksi (kg ve m2)
X7	Genetik Diyabet Yatkınlık	Genetik olarak Diyabet hastalığına yatkınlık durumu
X8	Yaş	Kişinin yaşı (yıl)
Y	Sonuç	Sınıf Değişkeni (0-1)

2.2. Veri Seti Önişleme

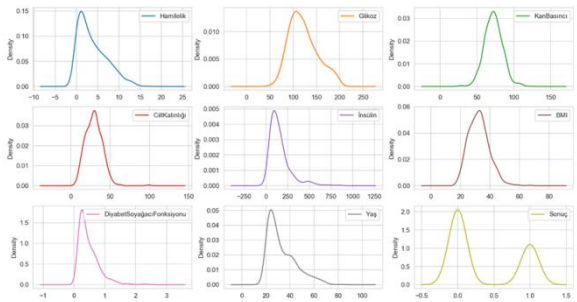
Verileri seti üzerinde yapılacak olan ön işlenme, çalışma kapsamında kullanılan makine öğrenmesi algoritmalarının sınıflandırma performans değerlerini artırmak ve en iyi sonucu elde etmek için uygulanan önemli adımlardan biridir. Makine öğrenmesi yöntemlerinde kullanılan algoritmanın performansı, kullanılan veriler arasındaki korelasyona, veri seti içerisinde eksik veya aykırı değere bağlı olarak değişmektedir. Bu çalışmada kullanılan veri seti düzenlenme işlemi üç aşamadan oluşmaktadır.

1) NaN yani eksik değerlerin doldurulması.

2) Özellik Ölçeklendirme (Feature Scaling)

3) Yeniden Örnekleme (Veri Seti Dengeleme)

Kullanılan veriler içerisinde eksik değerler (NaN) bulunmamasına rağmen Tablo 2 'ye baktığımızda Glikoz, İnsülin, Vücut Kitle İndeksi, Kan Basıncı ve Deri Kalınlığı gibi özelliklerinin min () değerinin "0" olduğu görülmektedir. İnsan vücudu ve anatomisi göz önüne alındığında ve düşünüldüğünde bu değerlerin "0" olamayacağı, aslında bu değerlerin eksik değerler olduğu görülmektedir. Eksik değerlerin doldurulması için kullanılan çeşitli yöntemler bulunmaktadır. Bunlardan biri de eksik değerlerin ***Çarpıklık*** (Skewness) durumuna göre değiştirilmesidir. Veri setinde bulunan her bir özelliğe ait yoğunluk grafikleri Şekil 2'de gösterilmiştir..



Şekil 2.Veri yoğunluk grafiği.

Çarpıklık bir dağılımın asimetri dağılımını, yani verilerin dağılımının simetrik olmama derecesini ölçer. Diğer bir ifade ile normal dağılımdan sapma miktarı hakkında bilgi verir. Bu çalışmada, pozitif çarpıklığa sahip özniteliklerin eksik değerleri o sütunun medyanı yani ortanca değeriyle (median), normal dağılıma ait özniteliklerin kayıp değerleri ise o sütunun ortalama (mean) değeriyle değiştirilmiştir. Tablo 3'te belirtilen özelliklere ait çarpıklık değerleri verilmiştir.

Tablo 3. Belirtilen özelliklerin çarpıklık değeri.

Öznitelik	Çarpıklık
Glikoz	0.5309
Kan Basıncı	0.1341
Cilt Kalınlığı	0.6906
İnsülin	2.166
Vücut kitle indeksi	0.5939

Glikoz, Kan Basıncı, Cilt Kalınlığı ve Vücut Kitle İndeksi gibi sütunlar o kadar çarpık değildir. Bu sütunlar için boş

değerler ortalama ile İnsülin sütunu çarpıklığın etkisinden dolayı medyan değeri olacak şekilde değiştirilmiştir.

Veri seti üzerinde yapılacak olan ikinci işlem veri (özellik) ölçeklemedir (Feature Scaling). Özellik Ölçeklendirme, bir veri kümesinde bulunan özelliklerin aralığını normalleştirme işlemidir. Makine öğrenmesinde kullanılan, özellikle uzaklık temelli algoritmaların performansını etkilemekle birlikte Gradyan Mesafesi (Gradient Distance) kullanılan algoritmaların da hız performansını etkilemektedir. Uzaklık temelli algoritmalar, benzerlikleri belirtmek veya bulmak için noktalar arasında bulunan mesafeyi kullanır. Daha büyük büyüklüğe sahip özellikler oluşturulan model tarafından daha yüksek ağırlıklı olarak belirecektir veya derecelendirilecektir ve oluşturulan model bir özelliğe aşırı derecede bağımlı olacaktır. Oluşturulan modelde tüm özelliklerin tahmin sonucuna eşit olarak katkı sağlaması için özellik ölçeklendirme uygulamamız önemlidir. Bu çalışma da standart ölçülendirme (Standart Scaler) yöntemi kullanılmıştır [5].

Standart ölçülendirme, bir veri kümesinde bulunan özellikleri değerleri benzer ölçeği paylaşacak şekilde özellikleri dönüştürme işlemidir. Veri setinde bulunan değişkenleri ortalaması ‘0’ standart sapması ‘1’ olan bir dağılıma çevirir. Herhangi bir sütunda bulunan x özelliğine sahip sütunun standart ölçeklendirme yöntemi Denklem 1’ de gösterilmiştir.

$$z = \frac{x-\mu}{\sigma} \qquad 1$$

$$\sigma = standart\ sapma$$

$$\mu = ortalama$$

2.3. Veri Seti Yeniden Örnekleme

Yeniden örnekleme, sınıf dağılımının eşit olmadığı durumlarda yapılan eğitim veri kümesine örnek eklemek veya çıkarmak için tasarlanmış, veri seti dengeleme işlemidir. Veri Seti dengeleme kavramı yani Dengesiz Veri Seti en basit haliyle bir grupta bulunan gözlem sayısının diğer gruba kıyasla daha az olması olarak tanımlanmaktadır.

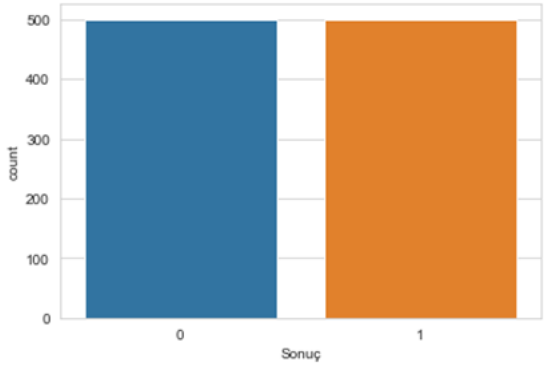
Makine öğrenmesinde kullanılan algoritmaların performans ölçümünü belirleyen etkenlerden biri de veri içerisinde eşit veya eşite yakın sayı sınıf etiket örneğinin bulunmasıdır. Veri Seti Dengelemek için farklı yaklaşım ve teknikler kullanılmaktadır. Bu çalışmada sınıf dengesizlik probleminin sınıflandırma üzerindeki olumsuz etkisini azaltmak için Sentetik Azınlık Aşırı-Örnekleme (SMOTE) tekniği kullanılmıştır [7-8]. Kullanılan algoritmanın asıl amacı azınlık sınıfında bulunan verilerin çoğunluk sınıfında bulunan veri miktarına yaklaştırılarak çoğaltılmasıdır. Yani örnekler oluşturmak için enterpolasyon tekniği kullanan sentetik aşırı örnekleme (over sampling) tekniğidir

SMOTE, en yakın komşu algoritması (k-NN) fikrine dayanır ve sentetik bir veri örneğinin orijinal ve en yakın komşulardan biri arasında enterpolasyon yapılabileceğini varsayar. SMOTE algoritması, azınlık sınıfından her veri örneğinin komşu ortamını hesaplar, komşularından birini rastgele seçer ve her örnek ile seçilen en yakın komşu arasındaki verilerin interpolasyonu yoluyla sentetik veri yapar. Yapılacak sentetik veri örneklerinin sayısı orijinal veri kümesinin boyutundan küçük olduğunda, algoritma rastgele seçilir ve sentetik veri örnekleri oluşturmak için orijinal bir veri örneği kullanılır. Tersine, yapılacak sentetik veri örneklerinin sayısı orijinal veri kümesinin boyutundan büyük olduğunda, algoritma önceden belirlenmiş aşırı örnekleme oranını kullanarak yinelemeli olarak sentetik veri örnekleri oluşturur [8].

Sentetik Örneklerin oluşturulması Denklem 2’ de gösterildiği gibi kısaca incelenen özellik vektörü E_i ile en yakın komşusu arasındaki alınır, daha sonra bu fark 0 ile 1 arasında rastgele bir sayı δ ile çarpılır. Sonuç incelenen özellik vektörüne eklenir ve yeni örnek oluşturulmuş olur.

$$E_{yeni} = E_i + \left(E_i - E_j\right)\delta \qquad 2$$

Çalışmama kapsamında kullanılan veri setinde bulunan sınıf etiketlerine baktığımızda veri setinde bulunan 768 kayıttan 268’i diyabet hastası, geriye kalan 500’ü diyabet hastası olmayan kişilere aittir. Verilerin %70 i eğitim, %30 kısmı test olarak ayrılmıştır. Kullanılacak sınıflandırma algoritmalarında 400 diyabet hastası olmayan sınıftan, 214 diyabet hastası olan grup olarak ayrılmıştır. SMOTE tekniği kullanılarak veri seti 500 pozitif, 500 negatif olarak düzenlenmiştir. Şekil 3 yeniden örnekleme işlemi yapıldıktan sonra elde edilen veri sayılarını göstermektedir.



Şekil 3.Veri seti sayısı.

Tablo 2. Veri kümesi değerleri.

	Gebelik	Glikoz	Tansiyon	Deri Kalınlığı	İnsülin	Vücut Kitle İndeksi	Diyabet geçmişi	Yaş	Sonuç
Ortalama	3.84	120.89	69.10	20.53	79.79	31.99	0.47	33.24	0.34
Standart Sapma	3.36	31.97	19.35	15.95	115.24	7.88	0.33	11.76	0.47
Minimum Değer	0.00	0.00	0.00	0.00	0.00	0.00	0.078	21.00	0.00
25%	1.00	99.00	62.00	0.00	0.00	27.30	0.24	24.00	0.00
50%	3.00	117.00	72.00	23.00	30.50	32.00	0.37	29.00	0.00
75%	6.0	140.25	80.00	32.00	127.25	36.60	0.62	41.00	1.00
Maximum Değer	17.00	199.00	122.00	99.00	846.00	67.10	2.42	81.00	1.00

3. Sınıflandırma

Bu çalışma kapsamında kullanılan sınıflandırma algoritmaları hakkında kısaca bilgi verilmiştir.

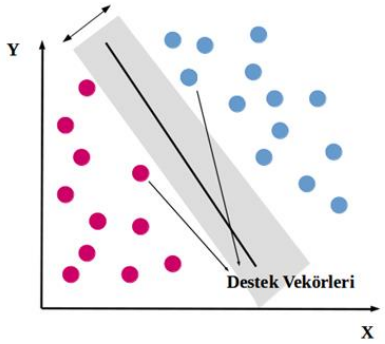
3.1. Naive Bayes Algoritması

Gausssan Naif Bayes sınıflandırıcısı veya genellikle sadece Naif Bayes olarak adlandırılan sınıflandırma algoritması, Bayes teoremine dayalı basit bir olasılık tabanlı sınıflandırma yöntemidir. Naif Bayes yöntemi tahmine dayalı modelleme yapmak için basit bunun yanında güçlü bir algoritmadır. Bu yüzden özellikle sinyal ve görüntü işleme alanlarında en çok kullanılan sınıflandırma, tahmin algoritmalarında biridir. Bu algoritmada bir sınıftaki belirli özelliklerin varlığının diğer herhangi bir özellik ile ilgisi olmadığı varsayılır [9].

3.2. Destek Vektör Makinası (DVM)

Destek vektör makineleri sınıflandırma ve regresyon için yaygın olarak kullanılan denetimli makine öğrenmesi tekniklerinden biridir ve Vapnik tarafından geliştirilmiştir [10]. İstatistiksel öğrenme teorisi ve yapısal risk minimizasyonu teknikleri üzerine kurulmuş olmasından dolayı teorik altyapısı güçlü makine öğrenmesi tabanlı örüntü sınıflandırma tekniğidir.

Destek vektör makinelerinin temel amacı eğitim verilerini bilinen sınıf etiketleriyle ayırabilen çok boyutlu uzayda bir fonksiyon bulmaktır. Sınıf etiketleri genellikle pozitif ve negatif olarak adlandırılan veri seti içerisinde Şekil 4 'te gösterildiği gibi en uygun ayıracı yani hiper düzlemi bulmaktır [11].



Şekil 4. Destek Vektör Makinası.

4. Sonuç

Çalışma kapsamında yapılan analizlerde Python kodlama dili kullanılmıştır. Diyabet hastalığının teşhis edilmesi için Gaussian Naive Bayes ve DVM makine öğrenmesi sınıflandırma algoritmaları kullanılmıştır. Makine öğrenmesinin en önemli parçalarından biri de sınıflandırma işlemidir. Sınıflandırma işleminde elimizde bulunan veriler eğitim seti ve test seti olmak üzere iki ana aşamaya ayrılır. Oluşturulan model eğitim seti kullanılarak, model performansı test seti kullanılarak yapılır. Buradan anlaşılacağı gibi veri seti makine öğrenmesi algoritmalarında çok önemlidir. Oluşturulan modelin değerlendirme kriterleri Tablo 4'te açıklamaları ve formülleri verilmiş olan Doğruluk Oranı (Accuracy Rate), Kesinlik (Precision), Duyarlılık (Recall) ve F1-Skore (F1 Score) değerleri kullanılarak hesaplanmıştır. Performans ölçüm değerlerin hesaplanması için Tablo 5 'te gösterilen karmaşıklık (hata) matrisi kullanarak yapılmıştır.

Tablo 4. Model değerlendirme kriterleri.

Değerlendirme Kriter	Açıklama	Formül
Doğruluk Oranı	Oluştular modelin hedef sınıfları tahmin başarısı.	(DP+DN)/ N
Kesinlik (Precision)	Oluşturulan model de sonucun ne kadar doğru olduğunu gösterir.	DP/(DP+YP)
Duyarlılık (Recall)	Oluşturulan model de doğru örnekleri bulma yeteneğini gösterir.	DN/(YP+DN)
F1- Skore (F1 Score)	Kesinlik ve duyarlılığın harmonik ortalamasıdır.	2*(Kesinlik*Duyarlılık)/ (Kesinlik+Duyarlılık)/

Tablo 5.Hata matrisi.

	Pozitif	**Negatif**
Pozitif	DP (Doğru Pozitif / TP)	YN (Yanlış Negatif/ FN)
Negatif	YP (Yanlış Pozitif / FP)	DN (Doğru Negatif / TN)

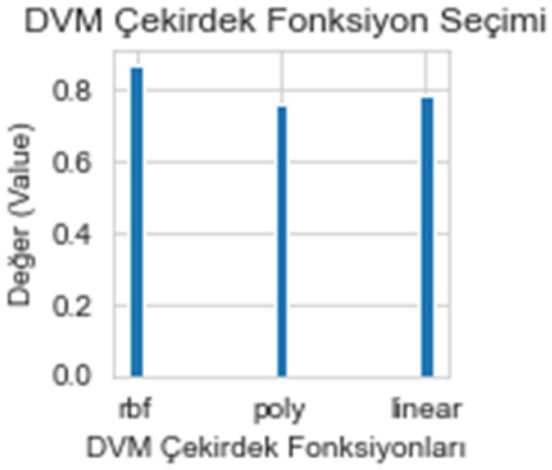
***** DP (TP): Gerçek sınıfın değeri pozitiftir (yani diyabet hastası) ve kullanılan yöntemle pozitif (diyabet hastası) olarak tahmin edilmiştir.

* DN (TN): Gerçek sınıfın değeri negatiftir (diyabet hastası olmayan) ve kullanılan yöntemle negatif (diyabet hastası olmayan) olarak tahmin edilmiştir.

* YN (FN): Gerçek sınıfın değeri pozitiftir (yani diyabet hastası) fakat kullanılan yöntemle negatif (diyabet hastası olmayan) olarak tahmin edilmiştir.

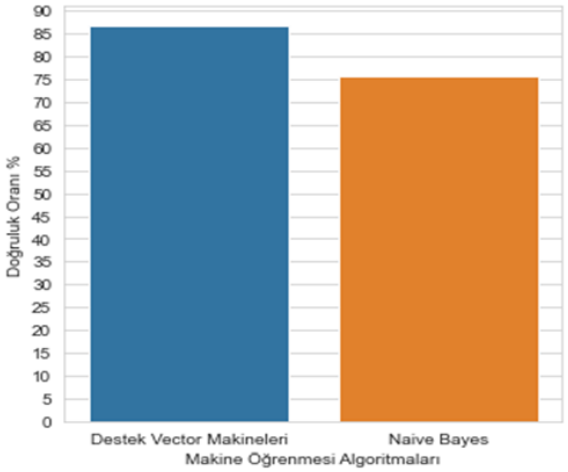
* YP (FP): Gerçek sınıfın değeri negatiftir (yani diyabet hastası olmayan) fakat kullanılan yöntemle pozitif (diyabet hastası) olarak tahmin edilmiştir.

DVM algoritması kullanılırken Raibal Basis, Siomoid ve Linear çekirdek fonksiyonlarının performansları Şekil 5'te belirtilmiştir. Buna göre en iyi performans sağlayan Radial Basis çekirdek fonksiyonu kullanılmıştır.



Şekil 5. DVM algoritması çekirdek fonksiyon seçimi.

Kullanılan algoritmaların performans ölçütleri kesinlik, duyarlılık, doğruluk oranı ve F1 score değerleri karşılaştırmalı olarak Tablo 6'da verilmiştir. Aynı zamanda kullanılan algoritmaların Doğruluk Oranına göre bar grafiği Şekil 6'da gösterilmiştir. Buna göre diyabet hastalığının sınıflandırılmasında en iyi sonucu %88 doğruluk oranı ile DVM algoritmasının verdiği görülmektedir.



Şekil 6. Naif Bayes ve DVM doğruluk oranı (%).

DVM algoritması sadece doğruluk oranı olarak değil aynı zamanda kesinlik ve duyarlılık için de yüksek sonuçlar vermiştir.

	Doğruluk Oranı	**Kesinlik**	**Duyarlılık**	**F1-Score**
Naif Bayes	%77	%76	%77	%76
DVM	%88	%87	%81	%87

Kaynakça

[1].https://www.who.int/health-topics/diabetes#tab=tab_1,21 Mayıs 2021 tarihinde alındı. (Erişim Tarihi: 31.05.2021).

[2].Özlüer Başer, B. , Yangın, M. & Sarıdaş, E. S. (2021). Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 25(1), 112-120. DOI: 10.19113/sdufenbed.842460.

[3].Srinivasa R, Yashashwini, Shubham janakatti, Venkatesh K B, Yaswanth S P. (2020). Prediction of Diabetes using Machine Learning. International Journal of Advanced Science and Technology, 29(06), 7593 - 7601.

[4].Kayaer, K., & Yıldırım, T. (2003). MEDICAL DIAGNOSIS ON PIMA INDIAN DIABETES USING GENERAL REGRESSION NEURAL NETWORKS. Proceedings of the International Conference on Artificial Neural Networks and Neural Information Processing, 2003.

[5].Bilgin, G. (2021). Makine Öğrenmesi Algoritmaları Kullanarak Erken Dönemde Diyabet Hastalığı Riskinin Araştırılması . Journal of Intelligent Systems: Theory and Applications , 4 (1) , 55-64 . DOI: 10.38016/jista.877292.

[6].Maniruzzaman, M., Rahman, M.J., Ahammed, B. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health Inf Sci Syst. 8, 7. https://doi.org/10.1007/s13755-019-0095-z.

[7]. Turhan, S., Yüksel, Ö., Şarer Yürekli, B. P., Suner, A., Doğu E. (2020) . Sınıf Dengesizliği Varlığında Hastalık Tanısı için Kolektif Öğrenme Yöntemlerinin Karşılaştırılması: Diyabet Tanısı Örneği. Türkiye Klinikleri Biyoistatistik Dergisi. 12 (1), 16-26. DOI: 10.5336/biostatic.2019-66816.

[8]. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. (2002. SMOTE). synthetic minority over-sampling technique. J. Artif. Int. Res. 16, 1 , 321–357

[9]. Cover, T. M. (1965). Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans. Electron. Comput. 14 (3), 326–334.

[10]. Cortes, C. and Vapnik, V. (1995) Support-Vector Networks. Machine Learning, 20, 273-297. http://dx.doi.org/10.1007/BF00994018.

[11]. Qi, Z., Tian, Y., & Shi, Y. (2013). Robust twin support vector machine for pattern classification. Pattern Recognit., 46, 305-316