GUFBD / *GUJS* (2022) 12(1): 211-226
DOI: 10.17714/gumusfenbil.820882 Araştırma Makalesi / *Research Article*

Diyabet tanısının tahminlenmesinde denetimli makine öğrenme algoritmalarının performans karşılaştırması

Performance evaluation of supervised machine learning algorithms for predicting diabetes mellitus

Yüksel ÖZKAN[1,a], Banu SARER YÜREKLİ[2,b], Aslı SUNER*[1,c]
[1] *Ege Üniversitesi, Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, 35100, İzmir*
[2] *Ege Üniversitesi, Tıp Fakültesi, Endokrinoloji Bilim Dalı, 35100, İzmir*

• Geliş tarihi / *Received*: 04.10.2020 • Düzeltilerek geliş tarihi / *Received in revised form*: 21.11.2021 • Kabul tarihi / *Accepted*: 06.12.2021

Öz
Hastalık tanısının doğru sınıflandırılmasında, hangi değişkenlerin analize alınacağı ve sonuçların nasıl değerlendirileceği klinik karar verme sürecinin yanı sıra istatistiksel yaklaşımda da doğru bir şekilde tanımlanmalıdır. Bu çalışmada en iyi sınıflandırma performansına sahip algoritmaya iki farklı yaklaşımla karar verilmesi amaçlanmıştır. Kullanılan veri seti, Haziran–Eylül 2013 arasında bir devlet hastanesinin endokrinoloji polikliniğine gelen yaşı 18 ve üstü olan toplam 232 hastadan elde edilmiştir. Diyabet tanısının sınıflandırılması için iki farklı yaklaşım kullanılmıştır. İlk yaklaşımda çokterimli lojistik regresyon yönteminde istatistiksel olarak anlamlı bulunan 18 değişken, ikinci yaklaşımda ise endokrinoloji uzmanı tarafından belirlenen ve klinik olarak önemli bulunan 21 değişkenle modeller kurulmuştur. Diyabet tanısı, denetimli makine öğrenme algoritmalarından Naïve Bayes, Bayes ağları, rastgele orman, karar ağaçları, destek vektör makinaları, k-en yakın komşuluk, yapay sinir ağları ve çokterimli lojistik regresyon yöntemleri ile sınıflandırılmıştır. Model performansları, doğrulukları, Kappa istatistikleri, ortalama mutlak hataları, hata kareler ortalamalarının karekökleri, göreceli mutlak hataları, duyarlılıkları, seçicilikleri, kesinlikleri, F-ölçütleri, Matthews korelasyon katsayıları, ROC eğrileri ve Youden indeksleri kriterlerine göre karşılaştırılmıştır. Model performanslarının test edilmesinde 10-katlı çapraz geçerlilik yöntemi uygulanmış, her algoritmanın çalışma süreleri hesaplanmıştır. Tüm analizler, WEKA 3.8.2 ve R Studio 1.1.383 ile yapılmıştır. Genel anlamda en iyi performansa sahip algoritma, rastgele orman algoritması olarak belirlenmiş, model doğrulukları sırasıyla %84.48 ve %81.90 olarak bulunmuştur. Diyabet hastalığının tanısının konulmasında, doğru sınıflandırma yapabilen modelin seçiminde klinik anlamlılığın yanı sıra istatistiksel anlamlılığa da önem verilmelidir.

Anahtar kelimeler: Denetimli öğrenme, Diyabet tanısı, Makina öğrenme algoritmaları, Prediyabet, Sınıflandırma

Abstract
In correct classification of disease diagnosis which variables are analyzed and how results are evaluated should be correctly defined in clinical decision making process as well as in statistical approach. It is aimed to determine the algorithm which has the best classification performance by using two different approaches in this study. The data set was obtained from 232 patients aged≥18 who were admitted to endocrinology outpatient clinic of a public hospital between June-September, 2013. Two different approaches were used to classify diagnosis of diabetes. In the first approach, 18 variables which were found statistically significant in multinomial logistic regression method were utilized; in the second approach, all models were built with 21 clinically significant variables which were determined by expert endocrinologist. Diabetes was classified with supervised machine learning methods; including Naïve Bayes, Bayes network, random forest, decision trees, support vector machine, k-nearest neighbors, artificial neural network and multinomial logistic regression. The performance of models was evaluated with accuracy, Kappa statistics, mean absolute error, root mean squared error, relative absolute error, sensitivity, specificity, precision, F-measure, Matthews correlation coefficient, ROC curve and Youden index. 10-folds cross-validation method was applied to test performance of models; runtimes of each algorithm were calculated. Analyses were performed with WEKA 3.8.2 and R Studio 1.1.383. Generally, random forest algorithm had the best performance with accuracy 84.48% and 81.90%, respectively. Clinical significance should be emphasized as well as statistical significance when choosing correct classification model for diagnosis of diabetes.

Keywords*: Supervised learning, Diabetes mellitus diagnosis, Machine learning algorithms, Prediabetes, Classification*

*[c] Aslı SUNER; asli.suner@ege.edu.tr, Tel: (0232) 390 19 85, orcid.org/0000-0002-6872-9901
[a] orcid.org/0000-0003-0534-1173 [b]orcid.org/0000-0003-1809-2655

1. Giriş
1. Introduction

Tüm dünyada diyabetli kişi sayısı, tahmin edilenden daha hızlı bir şekilde artmaktadır. Diyabetin önlenmesi ve yönetilmesi için yapılacak çalışmaların artmasında, birinci basamak sağlık hizmetlerinde hizmet veren hekimlere büyük görev düşmektedir (Goldenberg & Punthakee, 2013). Uluslararası Diyabet Federasyonu (IDF)'nun 2021 yılı tahminlerine göre, 10 yetişkinden birinin (20-79 yaş) diyabet hastası olması (537 milyon) beklenmektedir (International Diabetes Federation, 2021). Diyabet tanısının sınıflandırılmasında glukoz tolerans durumuna göre normal, prediyabet (Pre) ve diyabetes mellitus (DM) olmak üzere üç ayrı gruptan bahsedilebilmektedir. Diyabet hastalığı, insülin salınımının, insülin etkisinin veya her iki faktörün birlikte oluşturduğu bozukluk nedeniyle yüksek kan şekeri (*hiperglisemi*) sonucu ortaya çıkan kronik metabolik bir hastalık olarak tanımlanmaktadır (Egan & Dinneen, 2014). Bunun yanı sıra, "gizli şeker" olarak adlandırılan prediyabet durumunda ise, kan şekeri düzeyi normal değerden yüksek olmasına karşın, diyabet tanısı koyacak kadar yeterli yüksekliğe sahip olmamaktadır (American Diabetes Association, 2014). Prediyabet tanısı konulan kişilerde, tip 2 diyabet ve kardiyovasküler hastalıkların gelişme riski daha fazladır (Bansal, 2015). Günümüzde oldukça yaygın olarak görülmekte olan diyabet hastalığının ülkelere maliyeti de göz önünde bulundurulduğunda, hastalığın erken teşhis edilmesi ve bir an önce uygun tedaviye başlanması ile hasta sonuçlarında iyileştirmeye gidilebilir ve bu hastalığa ilişkin ulusal harcamalar azaltabilir (Bilgin, 2021).

İnsan sağlığını bu derece tehdit eden diyabetin, erken tanı ve teşhisinin konulmasında kullanılacak sınıflandırma model yaklaşımları da önem kazanmaktadır. Bu anlamda, makine öğrenme yöntemleri, tıbbi tanı ve tedavinin önemli görevlerine yardımcı olmak amacıyla faydalı bilgileri sunmak için uygulanmaktadır. Bu yöntemler, kullanılan tıbbi verilerin farklı perspektiflerden analiz edilmesini ve yorumlanmasını sağlamaktadır. Diyabet tanısında farklı yaklaşımlarla makine öğrenme algoritmaları uygulanmıştır. Bu konuda, diyabet olan veya olmayan bireylerin tanımlanması için nicel kitle-sağlık ilişkisi (*quantitative population-health relationship-QPHR*) modelinin oluşturulmasında hematolojik parametreler ile glisemik durum arasındaki ilişki araştırılmıştır (Worachartcheewan, 2013). Çalışmada toplam 190 hastanın kan şekeri düzeylerine göre normal, prediyabet ve diyabet olmak üzere üç sınıf oluşturulmuştur. Makine öğrenme algoritmalarından destek vektör makinesi ve yapay sinir ağları algoritmaları kullanılarak glisemik durumu tahminlenmiştir. Her iki algoritmanın da glisemik durumu sınıflandırırken %98'den fazla doğruluğa sahip olduğu sonucuna varmışlardır.

Diyabet hastalığı alanında makine öğrenme algoritmaları ile uygulama yapılırken yaygın olarak "Pima Indian diabetes mellitus" olarak bilinen veri seti kullanılmakta ve bu çalışmalarda diyabet sınıfının tahmin edilmesi amaçlanmaktadır (Kaggle, 2018). ABD Ulusal Diyabet ve Sindirim ve Böbrek Hastalıkları Enstitüsü (National Institute of Diabetes and Digestive and Kidney Diseases) tarafından oluşturulan bu veri seti, Kuzey Amerika'da yaşayan Pima yerlilerinin genetik olarak diyabete yatkın olmaları ve bu grupta diyabet görülme olasılığının yüksek olması nedeniyle tercih edilmektedir. Bu veri setini kullanan Karegowda ve arkadaşları (2012), geliştirdikleri karma modelde karar ağacı C4.5 ve k-ortalamalı kümeleme yöntemleri bir arada kullanmışlar ve bu modelin doğru sınıflandırma oranını, karar ağacı C4.5 yönteminin tek başına kullanıldığı durumdan daha yüksek elde etmişlerdir (Karegowda, vd., 2012). Maniruzzaman ve arkadaşları (2017), bu veri setindeki verileri sınıflandırmak için doğrusal diskriminant analizi, kuadratik diskriminant analizi ve Naive Bayes gibi çeşitli sınıflandırma algoritmalarıyla Gaussian süreci tabanlı sınıflandırma algoritmasıyla doğruluk, duyarlılık, seçicilik, pozitif kestirim değeri, negatif kestirim değeri ve ROC kriterlerine göre model performansları karşılaştırmışlardır (Maniruzzaman vd., 2017). Toplam 768 hastayı diyabet hastası ve kontrol sınıfı olmak üzere sınıflandırmışlardır. Aynı veri seti ile uygulama yapan bir başka çalışmada, en iyi performans gösteren beş sınıflandırma algoritmasını karar ağaçları, k-en yakın komşuluk, Lojistik regresyon, Naïve Bayes ve Rastgele orman olarak belirlemişlerdir. Bulgularında, sağlık kuruluşuna gelen yeni bir hastanın diyabet hastası olma olasılığını %84,78 doğruluk ve 0.912 AUC değeri ile tahminlemişlerdir (Özlüer Başer vd., 2021). Pima veri setiyle uygulama yapan Walia ve arkadaşları (2018) J48 karar ağacı, çok katmanlı algılayıcı, Naïve Bayes ve PART algoritmalarını kullanırken; Joshi ve Shetty (2015) J28, Bayes yaklaşımı, rasgele orman, Naïve Bayes, rasgele ağaç, REP, CART, birleşmeli kural öğrenme ve k-en yakın komşuluk algoritmaları ile algoritmaların performanslarını kıyaslamışlardır (Walia vd., 2018; Joshi & Shetty, 2015). Bir başka çalışmada

ise, Pima veri setinde evrişimsel sinir ağı ve uzun kısa süreli bellek ağları modellerinin hibrit olarak kullanıldığı yaklaşımla %86,45 oranında sınıflandırma başarısı elde edilmiştir (Er & Işık, 2021). Cihan ve Coşkun (2021) ise çalışmalarında, k-en yakın komşuluk, Naïve Bayes, lojistik regresyon, karar ağacı, destek vektör makinesi, rassal orman ve yapay sinir ağı algoritmalarını kullanarak Pima veri serinde uygulama yapmışlar ve lojistik regresyon yönteminin en yüksek performans değerlerine sahip olduğunu belirtmişlerdir (Cihan & Coşkun, 2021).

Farklı hasta verileri ile çalışan araştırmacılardan Olivera ve arkadaşları (2017), tanısı konmamış diyabeti tahminlemek için farklı makine öğrenme algoritmalarının performansını karşılaştırmışlardır (Olivera vd., 2017). Bu algoritmalarından Naïve Bayes, lojistik regresyon, k-en yakın komşuluk, yapay sinir ağları ve rastgele orman algoritmalarıyla modeller kurulmuştur ve en iyi modeller yapay sinir ağları ve lojistik regresyon kullanılarak oluşturulmuştur. Böylelikle, tahmin modellerinin çoğu benzer sonuçlar vermiştir ve kolayca elde edilen klinik veriler yoluyla tanısı konmamış diyabete sahip olma olasılığı yüksek olan kişilerin belirlenmesini sağlamışlardır. Chen ve Pan (2018) ise, Wenzhou Medical Üniversitesi'nin 35669 hastaya ilişkin veri setindeki klinik test sonuçlarını kullanarak algoritmaların sınıflandırma performanslarını kıyaslamışlardır. Diyabet tanısında kullandıkları LogitBoost algoritmasının, Adaboost.M1 algoritmasından biraz daha yüksek sınıflandırma başarısına sahip olduğunu bulmuşlardır (Chen & Pan, 2018). Bir başka çalışmada ise, Bangladeş Sylhet Sylhet Diyabet Hastanesi'ndeki 520 hasta veri ile diyabet olma olasılığını tahmin etmek için karar ağaçları, destek vektör makinaları, çok katmanlı algılayıcı yapay sinir ağları, topluluk öğrenme algoritmaları, k-en yakın komşuluk, doğrusal ayrımcı analizi makine öğrenme algoritmaları kullanılarak %99,81 oranı ile en yüksek doğruluk k-en yakın komşuluk algoritmasıyla elde edilmiştir (Bilgin, 2021). Bu çalışmada k-en yakın komşuluk algoritmasıyla bir diyabet erken tanı kiti geliştirmiştir.

Bazı çalışmalar, sadece tek bir denetimli makine öğrenme algoritmaları kullanarak farklı yaklaşımlar için alternatif sunmaktadır. Yu ve arkadaşları (2010), yaygın olarak görülen hastalıkları sınıflandırmak için destek vektör makinesi algoritmasına dayanan bir alternatif yaklaşım sunmuşlardır (Yu vd., 2010). Diyabet ve prediyabete sahip hasta sınıflarını iki farklı sınıflandırma ile destek vektör makinesi algoritmasının doğru sınıflandırma performansını ROC kriteri ile değerlendirmişlerdir. İlk sınıflandırmada, diyabet tanısı konmuş ve konmamış hastalar ile prediyabet ve normal hastalar karşılaştırılırken, ikinci sınıflandırmada; diyabet tanısı konmamış ve prediyabet hastalar ile normal hastalar karşılaştırılmıştır. Bu yaklaşımlar sonucunda; birinci sınıflandırma yaklaşımı en yüksek ROC eğrisi değerine sahip olarak en iyi ayırt ediciliğe sahip yaklaşım olarak saptamıştır ve destek vektör makinesi modellemesi, diyabet ve prediyabet gibi hastalıkları olan bireyleri tahminlemek için ileriye yönelik umut verici bir sınıflandırma yaklaşımı olduğunu savunulmuştur. Bu tür yaklaşımlarla, ortak değişkenler kullanılarak diğer karmaşık hastalıklarında sınıflandırılabileceğine dikkat çekmişlerdir. Diyabet tanısının sınıflandırmasında denetimli makine öğrenme algoritmaları kullanılarak çeşitli araçlar geliştirilmiştir. Heikes ve arkadaşları (2008), tanısı konmamış diyabet veya prediyabet hastalarına yönelik risk altında olan kişilerde diyabet tanısını tahminlemeye yardımcı olabilecek ve hiçbir hesaplama gerektirmeyen basit bir hesaplama aracı geliştirmişlerdir (Heikes vd., 2008). "*Diabetes Risk Calculator (DRC)*" adını verdikleri bu hesaplama aracı, kişileri 14 farklı kategoriye ayırarak her bir kategori için bir bireyin düşük risk altında, tanı konmamış diyabet ve prediyabete göre kategorize ederek yüksek risk altında olma olasılıklarını rapor etmektedir. DRC aracı geliştirmek için lojistik regresyon ile sınıflandırma ve regresyon ağacı (*classification and regression tree–CART*) denetimli makine öğrenme algoritmalarından yararlanılarak diyabet tanısı konulmaktadır. Aracın geliştirilmesinde CART algoritmasının seçilmesinin nedeni olarak oluşturulan karar ağacının basit bir araca dönüştürülebiliyor olması ifade edilmiştir. Böylelikle, CART ağacındaki yollar izlenerek bireyin diyabet veya prediyabet riskini belirlemişlerdir. Daghistani ve Alshammari (2016), diyabetli ve diyabetli olmayan hastaların doğru sınıflandırılması için kendini organize edebilen haritalama, C4.5 ve rastgele orman algoritmalarının performanslarını hassaslık ve kesinlik kriterlerine göre karşılaştırmışlardır (Daghistani & Alshammari, 2016). Rastgele orman algoritması, en iyi sınıflandırma performansına sahip algoritma olarak saptanmıştır.

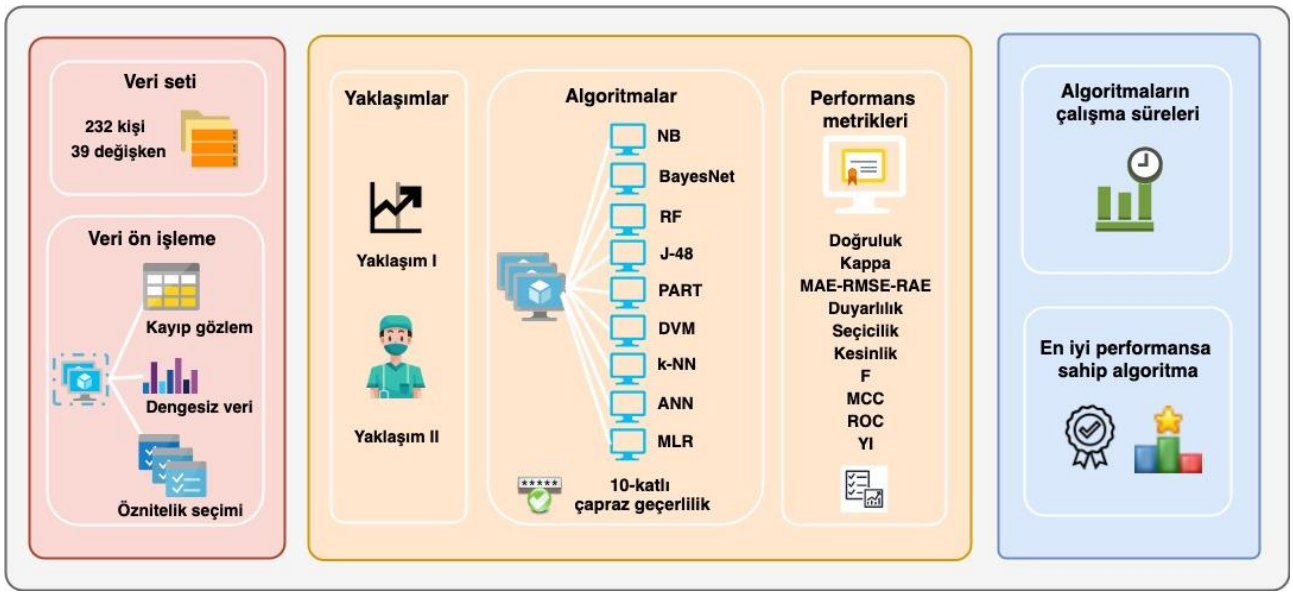
Bu çalışmada, diyabet tanısının sınıflandırılmasında denetimli makine öğrenme algoritmaları kullanılarak belirlenen kriterlerle model performansları karşılaştırılmış ve diyabet tanısını doğru sınıflandırabilen en iyi modele karar verilmesi amaçlanmıştır. En iyi modele karar

verirken hem istatistiksel önemlilik hem de klinik önemlilik göz önünde bulundurularak model performansları karşılaştırılmıştır.

2. Gereç ve yöntem
2. Material and method

Çalışmada kullanılan veri seti İzmir Bozkaya Eğitim ve Araştırma Hastanesi, Endokrinoloji ve Metabolizma Hastalıkları polikliniğinden Haziran – Eylül 2013 tarihleri arasında, 18 yaşından büyük hastalardan elde edilmiştir. Çalışmanın etik kurul onayı, 24.12.2013 tarihinde İzmir Bozkaya Eğitim ve Araştırma Hastanesi Klinik Araştırmalar Etik Kurulu'ndan alınmış ve Helsinki bildirgesi temel alınarak çalışmaya başlanmıştır. Toplam 232 hastanın yer aldığı çalışmada, 58 diyabeti ve prediyabeti bulunmayan birey, 32 prediyabet ve 142 diyabet hastası bulunmaktadır. Diyabet tanısının sınıflandırması için yedisi kategorik değişken olmak üzere toplam 39 bağımsız değişkenden yararlanılmıştır.

Diyabet tanısının konulmasında hangi değişkenin daha etkili ve önemli olduğunu belirlemek için iki farklı yaklaşım temel alınarak modeller kurulmuştur. Yaklaşım I'de sadece istatistiksel olarak anlamlı bulunan değişkenlerle modeller kurulmuştur. Öncelikle 39 değişken için tek tek çokterimli lojistik regresyon modeli kurulmuş, p-değeri anlamlı olan değişkenler seçilerek final modeller oluşturulmuştur. Yaklaşım II'de ise uzman hekim tarafından tüm değişkenler içerisinden en önemli değişkenler seçilip klinik olarak anlamlı olanların denetimli makina öğrenme algoritmaları performansları karşılaştırılmıştır. Tüm yaklaşımlarda değişkenlerin p-değerleri α = 0,05 anlam düzey ile karşılaştırılmıştır. Denetimli öğrenme algoritmalarından Naïve Bayes, Bayes ağları, rastgele orman, karar ağaçları (J–48 ve PART), k-en yakın komşu, yapay sinir ağları, destek vektör makinaları ve çokterimli lojistik regresyon algoritmaları kullanılarak doğruluk, Kappa istatistiği, ortalama mutlak hata, hata kareler ortalamasının karekökü, göreceli mutlak hata, duyarlılık, seçicilik, kesinlik, F-ölçütü, Mathews korelasyon katsayısı, ROC eğrisi ve Youden indeksi kriterlerine göre her iki yaklaşım için algoritmaların doğru sınıflandırma performansları karşılaştırılmıştır. Çalışmada kullanılan denetimli makine öğrenme algoritmaları aşağıda sırasıyla bahsedilmiştir. Son olarak, 10-katlı çapraz geçerlilik yöntemine göre eğitim ve test veri setleri oluşturulmuş, her algoritmanın çalışma süreleri hesaplanmıştır. Çalışmanın akış şeması aşağıdaki Şekil 1'de gösterilmiştir. Tüm istatistiksel analizlerde, WEKA 3.8.2 ve R Studio 1.1.383 kullanılmıştır.



Şekil 1. Çalışmanın akış şeması
Figure 1. *Flow chart of the study*

2.1. Naïve Bayes (NB)
2.1. Naïve Bayes (NB)

Naïve Bayes algoritması, Bayes koşullu olasılıklar teoremine dayanan bir sınıflandırma algoritmasıdır. Naïve olarak adlandırılmasının nedeni, sınıflandırma üzerinde etkisi olan değişkenlerin birbirinden bağımsız olduğu varsayımına dayanması ve Bayesçi sınıflandırma sayesinde yanlış sınıflandırma olasılığını en aza indirgemesidir. *p*, yüksek boyutlu veri setlerinde yoğunluk tahminin zor olduğu durumlarda

kullanılması önerilmektedir. Genel varsayımda her zaman doğru olmasa da basit anlamda bir sınıfın tahmini $G = j$ ve X_k sınıflandırma üzerinde etkisi olduğu düşünülen değişkenlerin bağımsız olduğu varsayıldığında,

$$f_k(X) = \prod_{k=1}^{p} f_{jk}(X_k) \quad (1)$$

bu tahminler, her bir sınıf-koşullu marjinal yoğunluklar f_{jk}, her biri için tek boyutlu Kernel yoğunluğu tahminleri kullanılarak ayrı ayrı tahmin edilebilmektedir (1). Bu aslında, Bayes teoreminin, marjinal yoğunlukları ifade ederken kullandığı tek değişkenli Gauss prosedürünün bir genellemesidir. Bu çalışmada, sürekli değişkenlerin Gauss dağılımına uygun şekilde dağıldığı varsayılmaktadır. Veri setindeki değişkenlerin dağılımlarının daha iyi tahminlenebilmesi için çekirdek tahmincisi algoritması kullanılmıştır. Eğer, X'in X_j bileşeni kesikli ise, uygun bir histogram tahmini kullanılarak değişken vektöründe değişken tiplerinden dolayı oluşacak sorunlar ortadan kaldırılmaktadır (Trevor vd., 2011). NB sınıflandırıcılar genellikle çok daha karmaşık algoritmalara kıyasla anlaşılması daha kolay ve sade bir sınıflandırma algoritması olmasının yanı sıra olasılıksal bilginin temsil edilmesinde, kullanılmasında ve öğrenilmesinde basit bir yaklaşım sunmaktadır. Fakat, özniteliklerin sınıflara göre bağımsız oldukları varsayıldığından, gereksiz özniteliklerin de modele eklenmesiyle öğrenme sürecinde sıkıntılar oluşabildiğinden, her algoritmada olduğu gibi NB algoritması da her veri setinde çok iyi sonuçlar veremeyebilmektedir.

2.2. Bayes ağları (BayesNet)
2.2. Bayesian Networks (BayesNet)

Bayes ağları (*Bayes Networks*) algoritması, Bayes teoremine dayanan veri setindeki değişkenler arasında var olan olasılıksal bağımlılıklar hakkında bilgi sağlamaktadır.

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \pi_i) \quad (2)$$

Bayes ağları, modeller oluşturmak için kullanılabilen olasılıksal grafik modeli türüdür. Bayes ağı iki bölümden oluşmaktadır. Bu bölümlerden biri yönlendirilmiş asiklik grafik şeklindeki kategorik bileşeni ve diğeri ise koşullu olasılık formundaki sürekli bileşenidir (2). Bayes ağları, mevcut durum ve bilgilere bağlı olarak çeşitli şekillerde oluşturabilmektedir. Özellikle, alanının uzmanlarıyla birlikte yüksek düzeyde bilgiler sentezlenerek elde edilen verilerden öğrenilerek Bayes ağları oluşturabilmektedir (Van Harmelen vd., 2008).

2.3. Rastgele orman (RF)
2.3. Random forest (RF)

Rastgele orman (*random forest*) algoritması, torbalama yönteminin gelişmiş bir şekli olarak kabul edilen ve hem sınıflandırma hem de regresyon algoritması olarak kullanılmaktadır. Algoritma ağaç sayısı ve her düğüm ayrılmasında rastgele seçilen bağımsız değişkenlerin sayısı olarak iki parametre üzerine kurulmaktadır (Breiman, 2001). Karar ağaçları oluşturulurken, orijinal veri setine ait örneklem sayısı kadar bootstrap yöntemi (yeniden yerine koyarak) ile örneklem oluşturulur. Rastgele orman algoritması aşağıdaki şekilde kurulmaktadır:

1. Bootstrap yöntemi ile elde edilen veri seti, eğitim ve test veri setleri olmak üzere ikiye ayrılmaktadır.

2. Eğitim veri setinden CART karar ağacı algoritmasıyla en geniş ağaç elde edilmekte ve karar ağacı budanmamaktadır. Gini indeksi yöntemiyle ağacın hangi değişkenle bölünmeye başlayacağı belirlenmektedir. Bu yöntem, her düğümde yeni bir dal oluşmayana kadar tekrar edilmektedir.

3. Her yaprak düğüm bir sınıfa atanarak test veri setiyle karar ağacı modeli kurulmaktadır.

4. 1. adımdan 3. adıma kadar tüm adımlar tekrar edilmektedir. Böylelikle, en basit şekilde rastgele orman algoritmasıyla model kurulmaktadır.

Algoritmada genel olarak genelleme hatası, parametre ayarlanması, değişken önemliliği, örneklemler arası uzaklık ve kayıp değer ataması gibi özellikler büyük öneme sahiptir. Karar ağacında bulunan değişkenlerin tahmin ediciliğini ölçmek için değişken önemliliği hesaplanmaktadır. Bu çalışmada, oluşturulacak ağaç sayısına bir kısıt getirilmeyerek budama yapılmamıştır. Sınıflandırma modelleri için varsayılan parametre değeri, bağımsız değişkenlerinin sayısının kareköküdür (aşağı yuvarlanmış sayısı). Bu çalışmada, $\sqrt{39} = 6{,}245 \cong 6$ olarak belirlenmiştir.

2.4. Karar ağaçları (J-48 ve PART)
2.4. Decision trees (J-48 and PART)

Karar ağacı tabanlı algoritmalardan biri olan J-48 algoritması, C4.5 algoritmasının açık kaynak kodlu bir Java uygulaması ve ID3 algoritmasının bir uzantısıdır (Goyal & Mehta, 2012). Sınıflandırma verilerinin doğru yönetilmesine ve yeni veriler hakkında tahminler yapılmasına yardımcı olmaktadır. Bir başka karar ağacı algoritması olan PART algoritması, her seferinde kısmi C4.5 algoritmasına dayanan budanmış karar oluşturarak en iyi yaprağı oluşturmaya çalışmaktadır. Bu algoritma, tekrar tekrar oluşturulan karar ağaçları ile birlikte böl ve fethet yöntemini kullanmasından dolayı esnek ve hızlı bir yapıya sahiptir. Temel olarak amaç, bütün bir karar ağacı oluşturmak yerine kısmi karar ağaçları oluşturmak olduğundan, ağaçları bir dizi örneklemlere ayırmaktadır. C4.5 algoritmasını kullanarak alt kümelere ayırdığı yapıyı, en küçükten başlayarak ortalama entropi sırasına göre genişlemektedir. Bu işlem alt kümeler yaprak haline gelen kadar devam etmekte ve iç düğümler ortaya çıkmaya başladığı anda budama işlemi yapılmaktadır. Böylelikle, düğümün daha iyi bir yaprak ile değiştirilip değiştirilmediğini kontrol edilmektedir. Bu çalışmada kullanılan karar ağacı algoritmalarında ağaçlarda budama yapılmamıştır.

2.5. Destek vektör makinaları (DVM)
2.5. Support Vector Machine (SVM)

Destek vektör makinaları algoritması, ikili sınıflandırma için geliştirilmiş, ancak zamanla hem çoklu sınıflandırma hem de regresyon modeller için de kullanılmaya başlanmıştır. DVM en başta sadece sürekli değişkenlerin analizi için kullanılmasına rağmen günümüzdeki haliyle kategorik değişken analizi için de kullanılmaktadır. Kategorik değişkenler, otomatik olarak sayısal değerlere dönüştürülerek hem kategorik hem de sürekli değişkenler normalize edilmektedir. Algoritmanın çalışma yaklaşımında; düzlem üzerinde iki sınıf arasındaki en uygun ayrımın sağlanması amaçlanmaktadır. Çakışan sınıflar durumunda, diskriminant marjininin veri noktalarındaki etkilerini azaltmak için aşağıya çekmesi, doğrusal olmayan durumda veri noktalarını etkili bir şekilde doğrusal olarak ayrılabilmesi için yüksek boyutlu uzaya yansıtılması (çekirdek yöntemi) ve problemin çözümü noktasında ikinci dereceden bir optimizasyon problemi olarak formüle edilmesi gibi temel yaklaşımları kullanmaktadır (Liv vd., 2014).

DVM'da farklı parametreler kullanılmaktadır. Karmaşıklık *(complexity)* parametresi, sınıfları ayırmak için kullanılacak çizginin ne kadar esnek olabileceğini kontrol etmektedir. 0 değeri, kenar boşluğunun ihlal edilmesine izin vermezken, varsayılan değer 1'dir. Diğer bir anahtar parametre ise, kullanılacak çekirdek türünün belirlenmesidir. En basit çekirdek, verileri düz bir çizgi veya hiperdüzlemle ayıran bir doğrusal çekirdektir *(linear kernel)*. Eğri ya da kıvrımlı bir çizgi kullanarak sınıflara ayıracak olan polinom çekirdeği *(polynomial kernel)*, polinom ne kadar yüksekse, o kadar zayıftır *(üs değeri)*. En popüler ve güçlü bir çekirdek olan radyal temel işlev çekirdeği *(radial basis function (RBF) kernel)*, sınıfları ayırmak için kapalı çokgenler ve karmaşık şekiller kullanmaktadır. Bu çalışmada varsayılan parametre olarak polinom çekirdek kullanılmıştır. Yöntemde girdi vektörleri yüksek boyutlu düzlemde doğrusal olmayan şekilde eşlendikten sonra bu düzlem üzerinde doğrusal bir karar yüzeyi oluşturularak öğrenme makinasının genellenebilirliği sağlanmaktadır.

2.6. K-en yakın komşuluk (k-NN)
2.6. K-nearest neighbours (k-NN)

K-en yakın komşuluk (*k-nearest neighbours*) algoritması, parametrik olmayan ve bellek tabanlı öğrenme (*memory based learning*) sınıflandırma algoritmasıdır. Bellek tabanlı öğrenme olmasının nedeni algoritmanın bir model öğrenmediği anlamına gelmektedir. Bunun yerine, tahminleme için bilgi olarak kullanılan eğitim örneklerini ezberlemeyi seçmektedir. Algoritma, bilinen bir sorgu noktasında, x_0, x_0'e en yakın k eğitim noktalarının $x_{(r)}$ $r = 1, 2, \ldots, k$ bulunarak k komşular arasında çoğunluk oyu kullanılarak sınıflandırılmasına dayanmaktadır. Benzerlik, iki veri noktası arasındaki bir uzaklık metriğine göre tanımlanmaktadır. En çok tercih edilen uzaklık ölçüsü öklid uzaklığı iken; Manhattan, Chebyshev ve Hamming uzaklıkları da çalışmalarda kullanılmaktadır. Öklid uzaklığının hesaplanması (3) ile gösterilmektedir:

$$d_{(i)} = \|x_{(i)} - x_0\| \quad (3)$$

Algoritmanın herhangi bir genelleme yapmak için eğitim veri setlerini kullanmıyor olması, diğer algoritmalara göre oldukça hızlı olmasını sağlamaktadır. Böylelikle tüm eğitim veri setini belleğinde tutmakta ve test aşamasında tüm eğitim verilerine ihtiyaç duymaktadır. Mahallenin büyüklüğü *(size of the neighborhood) k* parametresi ile ifade edilmektedir. Eğer k değeri 1 olarak belirlenirse, yanlılık düşük olmasına rağmen

varyans büyük olabilmektedir (Trevor vd., 2011). Aynı zamanda *k* değerinin 1 olması öngörülecek olan yeni modellerin (*patterns*) en yakın olan tek eğitim örneği kullanılarak tahminlerin yapılması anlamına gelmektedir. Veri setinin büyüklüğüne göre *k* değeri farklı değerler alabilmektedir. Bizim çalışmamızda hem sürekli hem de kategorik değişkenlerin bulunması nedeniyle uzaklık parametresi olarak Manhattan uzaklığı kullanılmıştır.

2.7. Yapay sinir ağları (ANN)
2.7. Artificial neural network (ANN)

Yapay sinir ağları *(artificial neural network)* algoritması, geniş bir sınıfın modellenmesini ve öğrenilmesini kapsayacak şekilde geliştirilmiştir. Tek gizli katman geri yayılım ağı veya tek katman algılayıcı olarak da adlandırılmaktadır. Doğrusal olmayan istatistiksel modellemede tercih edilen bir yöntem olmasının yanı sıra hem regresyon hem de sınıflandırma modellerinde kullanılmaktadır (Trevor vd., 2011). Çok değişkenli bir YSA, giriş değişkenlerinin doğrusal olmayan kombinasyonlarını kullanarak bir sistemin modellerini oluşturmak için tercih edilmektedir. YSA'nın birçok öğrenme kuralı olmasına rağmen, en çok kullanılan kurallar delta ve geri yayılımlı ağlar kurallarıdır. Sinir ağları, verileri haritalandırmak üzerine eğitilmekte ve girişlerden gelen bilgiler sinirler arasındaki ağırlıkları optimize etmek için ağlar üzerinden iletilmektedir. YSA, eğitim veri setindeki girdi ve çıktı değerlerini okuyarak tahmin edilen ile hedef değerler arasındaki farkı azaltmak için ağırlıklı bağlantıların değerini değiştirmektedir. Ağ, belirlenen doğruluk düzeyine ulaşıncaya kadar tahminlerdeki hatalar eğitim döngüsünde en aza indirgenmeye çalışılmaktadır.

2.8. Çokterimli lojistik regresyon (MLR)
2.8. Multinomial logistic regression (MLR)

Denetimli öğrenme yöntemlerinden olan çokterimli lojistik regresyon (*multinomial logistic regression*) algoritması, lojistik regresyon gibi bir sınıflandırma algoritmasıdır. Yüksek boyutlu veri yapılarında iyi çalışmaktadır (Böhning, 1992). $Y = (y_1, \dots, x_{k+1})^T$ gözlem vektörüne karşılık gelen p_j olasılığında çokterimli logit-modelde (4) ve (5) ile,

$$p_i = \frac{\exp\left(\pi^{(i)^T}\mathbf{x}\right)}{1+\sum_{j=1}^{k}\exp\left(\pi^{(j)^T}\mathbf{x}\right)} \qquad i = 1, \dots, k \tag{4}$$

$$p_{k+1} = \frac{1}{1+\sum_{j=1}^{k}\exp\left(\pi^{(j)^T}\mathbf{x}\right)} \tag{5}$$

$\mathbf{x} = (x_1, \dots, x_m)^T$: kovaryant vektörü ve π^i: i. yanıt kategorisine karşılık gelen parametre vektörüyle tanımlanmaktadır. Bu çalışmada, bir tür model düzeltme ya da ayarlama yaklaşımı olarak Ridge tahmincisi kullanılarak model oluşturulmuştur. Bu yöntem, model tarafından öğrenilen katsayıların değerlerini en aza indirerek eğitim sırasında modeli basitleştirmeyi amaçlamaktadır. Ridge parametresi, katsayıların genişliğini azaltmak için algoritmaya ne kadar basınç uygulanacağını tanımlanmakta ve 0'a çok yakın bir değer olarak alınmaktadır.

2.9. Algoritmaların performans karşılaştırma kriterleri
2.9. Performance comparison criteria of algorithms

Denetimli makina öğrenme algoritmalarının performans karşılaştırmasında birçok farklı kriter kullanılmaktadır. Bu çalışmada da her iki yaklaşım için ayrı ayrı en iyi performansı veren algoritmanın belirlenmesi için farklı performans karşılaştırma kriteri hesaplanmıştır.

2.9.1. Doğruluk
2.9.1. Accuracy

Bir testin doğruluğu, hasta ve sağlıklı bireyleri doğru olarak ayırt edebilme gücüdür (Baratloo vd., 2015). Tanı testinin doğruluğu hesaplanırken, tüm hasta ve sağlıklı bireyler için doğru pozitif ve doğru negatif oranı hesaplanmaktadır (Drobatz, 2009). Doğruluk değeri 0 ile 1 arasında değer almaktadır. (6) ile verilen gösterimde, DP: Doğru Pozitif; DN: Doğru Negatif; YP: Yanlış Pozitif ve YN: Yanlış Negatif olarak kısaltılmıştır.

$$\text{Doğruluk} = \frac{\text{DP}+\text{DN}}{\text{DP}+\text{DN}+\text{YP}+\text{YN}} \tag{6}$$

2.9.2. Kappa istatistiği
2.9.2. Kappa statistics

Gözlemciler arası güvenilirliği test etmek için sıklıkla kullanılan Kappa istatistiği, ilk olarak Cohen (1960) tarafından önerilen bir uyum ölçüsü kriteridir (Cohen, 1960). -1 ile 1 arasında değişen değerler alabilen Kappa istatistiğinin önceden belirlenmiş bir kabul edilebilirlik seviyesi olmamakla birlikte, 0,75 değerinden büyük olması mükemmel uyumu; 0,40 değerinden küçük olması zayıf uyumu ve 0,40 ile 0,75 arasındaki değerler alması ise kabul edilebilir uyumu göstermektedir. p_o, gözlenen uyum oranı ve p_e, beklenen uyum oranı olarak alındığında κ ile gösterilen Kappa

istatistiği aşağıdaki (7) ile belirtildiği şekilde hesaplanmaktadır:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (7)$$

2.9.3. MAE – RMSE – RAE kriterleri
2.9.3. MAE – RMSE – RAE criteria

Model performansları karşılaştırılırken, modele ait hatalardan hesaplanan hata mutlak ortalaması, hata kareler ortalamasının karekökü, hataların mutlak karekökü kriterleri kullanılmaktadır. MAE, gözlenen ve beklenen değerler arasındaki hatanın ortalama büyüklüğünü yönlerini dikkate almadan ölçerken, RMSE sadece hatanın ortalama büyüklüğünü ölçmektedir. Örneklem genişliği n birim olan modelin hataları $e_1, e_2, \ldots, e_n$ şeklinde ifade edilmektedir (Chai & Draxler, 2014). Buna göre MAE, RMSE ve RAE kriterleri aşağıdaki (8), (9) ve (10) ile hesaplanabilmektedir:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (8)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} \quad (9)$$

$$RAE = \frac{\sum_{i=1}^{n}|e_i|}{\sum_{i=1}^{n}|\bar{y} - y_i|} \quad (10)$$

2.9.4. Duyarlılık, seçicilik ve kesinlik
2.9.4. Sensitivity, specificity and precision

Bir testin duyarlılığı, bir bireyin gerçekten hasta olduğu bilindiğinde test sonucunda da hasta olduğu sonucuna varılmasıdır (11). Hasta bireylerin doğru olarak ayırt edilmesi için doğru pozitif oranı hesaplanmakta ve bu değer 0 ile 1 arasında değişen değerler almaktadır. Bir testin seçiciliği ise, bir bireyin gerçekten hasta olmadığı bilindiğinde, tanı testi sonucunun negatif çıkması olasılığı olarak tanımlanmaktadır (12). Sağlıklı bireylerin doğru ayırt edilmesi için, doğru negatif oranı hesaplanmakta ve bu değer 0 ile 1 arasında değişen değerler almaktadır (Baratloo vd. 2015). Kesinlik değeri, tanı testinin pozitif kestirim değeri veya pozitif kestirimlerin gerçekten pozitif olması durumda hesaplanmaktadır (13). Başka bir ifade ile tanı testi sonucu pozitif olan bir bireyin hasta olma olasılığı olarak tanımlanmaktadır.

$$\text{Duyarlılık} = \frac{DP}{DP + YN} \quad (11)$$

$$\text{Seçicilik} = \frac{DN}{DN + YP} \quad (12)$$

$$\text{Kesinlik} = \frac{DP}{DP + YP} \quad (13)$$

2.9.5. F-ölçütü
2.9.5. F-Measure

F-ölçütü, hassaslık ile kesinlik değerlerinin harmonik ortalaması şeklinde hesaplanmaktadır (14). Bu durumda, hem yanlış pozitif hem de yanlış negatif değerleri aynı anda dikkate alarak hesaplama yapmasından dolayı doğruluk olarak anlaşılması kolay değildir. Ancak düzensiz bir sınıf dağılımı varsa F-ölçütü değerinin doğrulukla birlikte değerlendirilmesi önerilmektedir. Eğer, pozitif ve yanlış negatifler benzer maliyetlere sahipse doğruluk değerine; yanlış pozitiflerin ve yanlış negatiflerin maliyeti farklıysa, F-ölçütü değerine bakmak daha faydalıdır (Hripcsak & Rothschild, 2005).

$$F = \frac{2*\left(\frac{DP}{DP+YN}\right)*\left(\frac{DP}{DP+YP}\right)}{\left(\frac{DP}{DP+FN}\right)+\left(\frac{DP}{DP+YP}\right)} \quad (14)$$

2.9.6. Matthews korelasyon katsayısı (MCC)
2.9.6. Matthews correlation coefficient (MCC)

İki veya daha fazla sınıflandırmada, sınıflandırma modelinin ya da fonksiyonun ne derecede performansa sahip olduğunu değerlendirmek için tercih edilen bir kriterdir (15). Matthews korelasyon katsayısı -1 ile 1 arasında değer almaktadır. -1 katsayısı, ters sınıflandırma yani tahmin ve gözlem değerleri arasındaki toplam uyuşmazlığı gösterirken; 0 katsayısı, ortalama düzeyde sınıflandırma performansına veya modelin rastgele tahminlerde iyi olmadığı anlamına gelmektedir. 1 katsayısı ise, mükemmel sınıflandırmayı veya tahminlemeyi belirtmektedir (Boughorbel vd., 2017).

$$MCC = \frac{(DP*DN) - (DP*YN)}{\sqrt{(DP+YP)*(YN+DN)*(YP+DN)*(DP+YN)}} \quad (15)$$

2.9.7. ROC eğrisi
2.9.7. ROC curve

ROC eğrisi, tanı testlerinin ve tahmin modellerinin değerlendirilmesi, uygun eşik değerinin ve testin ayırt ediciliğinin belirlenmesi, iki ya da daha fazla tanı testinin performanslarının karşılaştırılması gibi durumlarda kullanılan bir yöntemdir. Eğri altında kalan alan, hasta ve sağlıklı bireylerde tanı koymada ne kadar bir ayırt etme gücüne sahip olunduğunu göstermektedir. ROC eğrisi, uygun eşik değerinin belirlenmesinde, tüm olası eşik değerleri (c) için duyarlılığa (duyarlılık(c)) karşı $1 -$ seçicilik(c) değerlerine ait bir grafik vermektedir. Eğri altında kalan alan, 0 ile 1 arasında değer almakta ve kitlenin prevalansından etkilenmemektedir. Rastgele olarak bir tanı testinin

alabileceği eğri altında kalan alan değeri 0,5 iken; mükemmel bir tanı testinin doğruluğu 1 olarak ifade edilmektedir (Hajian-Tilaki, 2013).

2.9.8. Youden indeksi (YI)
2.9.8. Youden index (YI)

Youden indeksi, ROC eğrisi gibi tanı testinin etkinliğini ölçmede ve en uygun eşik değerinin belirlenmesinde kullanılan bir kriterdir (Fatima & Pasha, 2017). Tanısal doğruluğun en sık kullanılan ölçütü olan ROC eğrisine rağmen, Youden indeksi de tercih edilmektedir. Youden indeksi, 0 ile 1 arasında değer almaktadır (16). Hasta ve sağlıklı bireylerin tam olarak ayrılması durumunda $J = 1$; tam örtüşme durumunda $J = 0$ değerini vermektedir.

$$\mathrm{J} = \text{maksimum}\{\text{duyarlılık(c)} + \text{seçicilik(c)} - 1\} \quad (16)$$

2.9.9. k-katlı çapraz geçerlilik yöntemi
2.9.9. K-fold cross validation method

Algoritma modellerinin öğrenilmesi sürecinde eğitim ve test veri setlerinde bazı verilerin dışarıda tutulmasından dolayı birtakım sorunlar oluşturmaktadır. Eğitim verilerinin azaltılmasıyla, veri setindeki önemli eğilimlerin kaybolması riski altında olduğundan yanlı tahminler yapılabilmekte, dolayısıyla model hatası artmaktadır. Bu durumu önlemek için, modelin eğitilmesinde ve doğrulanmasında verinin çoğunluğundan yararlanılmasını sağlayan bir yöntem gerekmektedir. k-katlı çapraz doğrulama yöntemi, bu tür durumlar için tercih edilmektedir. Genel anlamda k-katlı çapraz doğrulama yönteminde ilk olarak ayarlama parametresi (λ) belirlenmekte, veriler rastgele olarak k alt kümeye bölünmektedir. Dışarıda tutma yöntemi (holdout method) k kez tekrarlanarak böylelikle, doğrulama veri seti olarak adlandırılan test veri seti için her k alt kümelerinden biri kullanılmaktadır. Geriye kalan k-1 alt kümeleri ise eğitim veri setini oluşturmak için bir araya getirilmektedir. Böylece, her veri noktasında k tane doğrulama veri seti ve k-1 tane eğitim veri seti bulunmaktadır. Modelin toplam etkinliğini elde etmek için tüm *k.* bölümlerinin çapraz doğrulama hata tahmininin ortalaması alınmaktadır (17)(18).

$$\mathrm{E_k}(\lambda) = \sum_{\text{i}\epsilon\text{k.bölüm}} \left(\mathrm{y_i} - \mathbf{x_i}\hat{\beta}^{-\mathrm{k}}(\lambda)\right)^2 \quad \mathrm{k} = 1, 2, \ldots, \mathrm{K} \quad (17)$$

$$\mathrm{CV}(\lambda) = \frac{1}{\mathrm{K}} \sum_{\mathrm{k}=1}^{\mathrm{K}} \mathrm{E_k}(\lambda) \quad (18)$$

Seçilen her λ değeri için, çapraz doğrulama hatasının tahminini CV(λ) en küçük yapan λ seçilmelidir. Uyum için verilerin çoğunluğu kullandığından, yanlılık önemli ölçüde azaltılmaktadır. Verilerin çoğunluğu test veri setinde kullanıldığı için varyans da önemli ölçüde azalmaktadır. Eğitim ve test veri setlerinin değiştirilmesi de bu yöntemin etkinliğini arttırmaktadır. Genellikle *k* değeri 5 veya 10 olarak tercih edildiği gibi, herhangi bir değer de alabilmektedir (Trevor vd., 2011). Bu çalışmada k değeri 10 olarak belirlenmiştir.

3. Bulgular
3. Results

Çalışmada, yer alan 39 bağımsız değişken ile diyabet tanısı sınıflandırılmıştır. Sürekli değişkenler; yaş, kemik morfojenik proteini-4 (BMP-4), matriks gla-proteini (MGP), noggin, çözünür lektin benzeri oksitlenmiş LDL reseptörü (sLOX-1), lipokalin, vücut kitle indeksi (VKI), ayak bileği-kol basınç indeksi (ABI), açlık plazma glukozu (APG), HbA1c, kreatinin, alanin transaminaz (ALT), alkalin fosfataz (ALP), gama glutamil transferaz (GGT), ortalama trombosit hacmi (MPV), eritrosit dağılım genişliği (RDW), hemoglobin (Hb), trombosit (PLT), ürik asit, toplam kolesterol, trigliserit (TG), yüksek yoğunluklu lipoprotein kolesterol (HDL), düşük yoğunluklu lipoprotein kolesterol (LDL), plazmanın aterojenik indeksi (AIP), C-reaktif protein (CRP), fibrinojen, paratiroid hormonu (PTH), tiroit uyarıcı hormonu (TSH), serbest triiodotironin (sT3), serbest tiroksin (sT4), kalsiyum (Ca) ve fosfor (P)'dur. Kategorik değişkenler ise, cinsiyet (kadın/erkek), sigara içme durumu (hiç/aktif/bırakmış), statin kullanımı (yok/var), asetil salisilik asit (ASA) kullanımı (yok/var), hipertansiyon (HT) (yok/var), koroner arter hastalığı (KAH) (yok/var) ve serebrovasküler hastalık (SVH) (yok/var)'tır.

Yaklaşım I'de, öncelikle 39 değişken için tek tek MLR modeli kurularak istatistiksel olarak anlamlı bulunan değişkenler seçilmiştir. Tablo 1'de tek tek incelendiğinde istatistiksel olarak anlamlı bulunan tüm değişkenler için kurulan modelde sırasıyla; cinsiyet, BMP-4, MGP, Noggin, sLOX-1, VKI, statin kullanımı, ASA kullanımı, HT, APG, HbA1c, ALT, MPV, toplam kolesterol, HDL, LDL, AIP ve Ca üzere toplam 18 değişkenin anlamlı olduğu görülmüştür. Bu değişkenlerle kurulan MLR modeli sonucunda diyabet ve prediyabet tanısı için anlamlı çıkan değişkenler Tablo 1'de yer almaktadır.

Tablo 1. Yaklaşım I için kurulan modeldeki değişkenlere ait p-değerleri
Table 1. *P-values of the variables in the model established for Approach I*

Değişken (18 adet)	Dm p-değeri	Pre p-değeri
Cinsiyet = Erkek	0.001*	0.001*
BMP - 4	0.428	0.296
MGP	0.067	0.016*
Noggin	0.582	0.100
sLOX-1	0.008*	0.001*
VKI	0.412	0.021*
Statin kullanımı	0.017*	0.723
ASA kullanımı	0.001*	0.001*
HT	0.004*	0.001*
APG	0.002*	0.001*
HbA1c	0.001*	0.844
ALT	0.086	0.261
MPV	0.056	0.254
Toplam kolesterol	0.001*	0.001*
HDL	0.001*	0.006*
LDL	0.001*	0.004*
AIP	0.001*	0.001*
Ca	0.053	0.459

Dm: Diyabet, Pre: Prediyabet, *p < 0.05

Yaklaşım II için hangi değişkenlerin modele alınacağına endokrinoloji uzmanı tarafından karar verilmiştir. Değişken sayısının fazla olduğu durumlarda, modeldeki her açıklayıcı değişken için en az 10 birey önerilmektedir (Alpar, 2011). Bu nedenle toplam veri sayısının 10'da birini aşmayacak şekilde, 21 değişkenle MLR modeli kurulmuştur. Tablo 2'de ikinci yaklaşım için modelde bulunan değişkenlerden sadece lipokalin değişkeninin diyabet ve prediyabet tanısında istatistiksel olarak anlamlı olmadığı sonucuna varılmıştır ($p > 0{,}05$).

Tablo 2. Yaklaşım II için kurulan modeldeki değişkenlere ait p-değerleri
Table 2. *P-values of the variables in the model established for Approach II*

Değişken (21 adet)	Dm p-değeri	Pre p-değeri
Cinsiyet = Erkek	< 0.001*	< 0.001*
Yaş	0.346*	0.022*
BMP-4	0.019*	0.005*
MGP	< 0.001*	0.005*
Noggin	0.003*	< 0.001*
sLOX-1	0.658*	< 0.001*
Lipokalin	0.174	0.707
VKI	< 0.001*	< 0.001*
Sigara içme = Aktif	< 0.001*	< 0.001*
Sigara içme = Bırakmış	< 0.001*	< 0.001*
HT	< 0.001*	< 0.001*
APG	< 0.001*	< 0.001*
HbA1c	< 0.001*	< 0.001*
Ürik asit	0.015*	0.071*
Toplam kolesterol	< 0.001*	< 0.001*
TG	0.014*	< 0.001*
HDL	0.000*	< 0.001*
LDL	0.003*	< 0.001*
KAH	< 0.001*	< 0.001*
SVH	< 0.001*	< 0.001*
CRP	< 0.001*	< 0.001*
Fibrinojen	0.059*	0.011*

Dm: Diyabet, Pre: Prediyabet, *p < 0.05

Her iki yaklaşımda da anlamlı bulunan değişkenler Tablo 3'te verilmiştir. Her iki yaklaşım için diyabet ya da prediyabetten en az birinde anlamlı olan değişkenler cinsiyet, MGP, sLOX-1, VKI, HT, APG, A1c, toplam kolesterol, HDL, LDL ve AIP değişkenleridir. Her iki yaklaşımda da hem diyabet hem de prediyabet tanısında anlamlı olan değişkenler ise cinsiyet, sLOX-1, HT, APG, toplam kolesterol ve LDL değişkenleri olarak bulunmuştur.

Tablo 3. Her iki yaklaşımda anlamlı olan değişkenler
Table 3. *Variables that were significant in both approaches*

		Cinsiyet	Yas	BMP-4	MGP	Noggin	sLOX-1	Linokalin	VKI	Sigara icme	Statin kullanımı	ASA kullanımı	HT	ABI	APG	A1c	Kreatinin	ALT	ALP	GGT	MPV	RDW	Hb	PLT	Ürik asit	Toplam kolesterol	TG	HDL	LDL	AIP	CAD	CVD	CRP	Fibrinojen	PTH	TSH	sT4	sT3	Ca	P
Y-I (18)	DM	X					X				X	X	X		X	X										X		X	X	X										
	Pre	X			X		X		X			X	X		X											X		X	X	X										
Y-II (21)	DM	X		X	X	X	X		X	X			X		X	X									X	X	X	X	X		X	X	X							
	Pre	X	X	X	X	X	X		X	X			X		X	X										X	X		X	X	X	X	X	X						

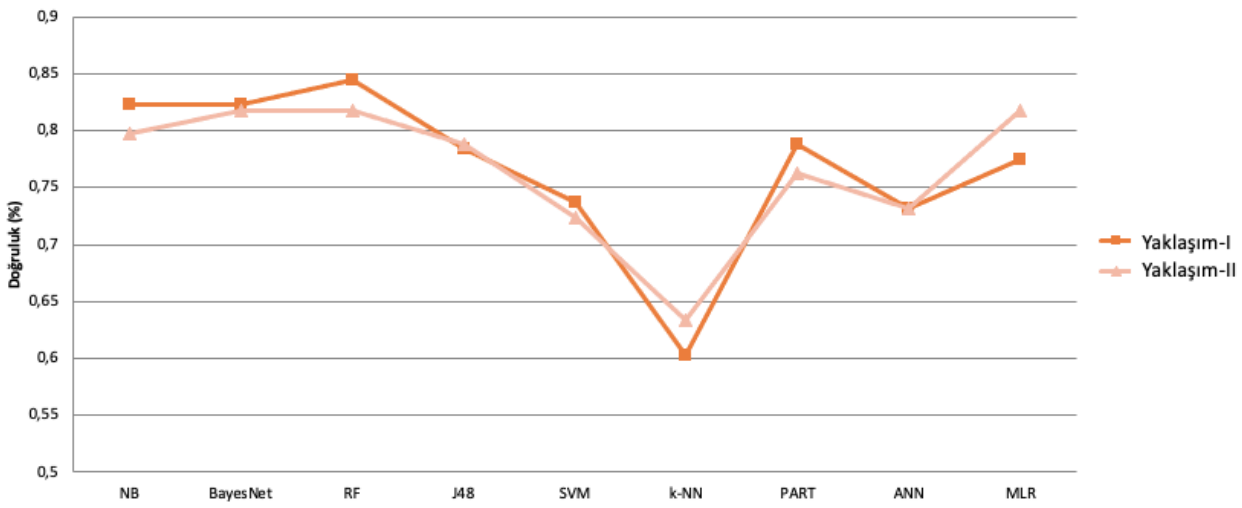
Dm: Diyabet, Pre: Prediyabet, Y-I: Yaklaşım I, Y-II: Yaklaşım II

Bir sonraki adımda, her bir yaklaşım için sınıflama amacıyla kullanılan tüm denetimli makina öğrenme algoritmalarının performans karşılaştırmaları Tablo 4'te özetlenmiştir. Yaklaşım I'de, 12 performans kriterinin 8'inde (%66,67) RF algoritması en iyi sınıflama performansına sahipken; Yaklaşım II'de sadece 5 kriterde en iyi performansı veren algoritmadır. Yaklaşım II'de Bayes ağları 12 kriterin 7'sinde (%58,33) en iyi sınıflama performansına sahip olan algoritma olmuştur. Her iki yaklaşım için model doğrulukları sırasıyla %84,48 ve %81,90 olarak bulunmuştur. Şekil 2'de de algoritmaların doğruluk oranlarına ilişkin değerler incelendiğinde, Yaklaşım I için en yüksek performans değerine sahip olan algoritma RF iken, Yaklaşım II'de BayesNET, RF ve MLR aynı değerlerle en yüksek performansa sahiptir. Veri seti, hasta ve sağlıklı durumunu gösteren sınıf değişkeni bakımından dengesiz veri özelliği taşıdığından, performans ölçütleri açısından değerlendirme yapılırken F-ölçütü, kesinlik ve duyarlılık ölçütleri öncelikli olarak dikkate alınmıştır. Her üç ölçüt için de Yaklaşım I'de RF algoritması, Yaklaşım II'de ise BayesNet en yüksek performans değerlerine sahip olan algoritmalardır. Genel anlamda bulgular incelendiğinde, performans ölçütleri açısından benzer sonuçlar elde edildiği Tablo 4'te görülebilmektedir.

Tablo 4. Her iki yaklaşım için denetimli makine öğrenme algoritmaları için performans karşılaştırması
Table 4. *Performance comparison for supervised machine learning algorithms for both approaches*

	Doğruluk	Kappa	MAE	RMSE	RAE	Duyarlılık	Seçicilik	Kesinlik	F-ölçütü	MCC	ROC	YI
Y-I	RF	RF	NB	RF	NB	RF	NB	RF	RF	RF	RF	NB
Y-II	BayesNet RF MLR	RF	J-48	RF	J-48	BayesNet RF MLR	BayesNet	BayesNet	BayesNet	BayesNet	RF	BayesNet

Y-I: Yaklaşım I, Y-II: Yaklaşım II



Şekil 2. Her iki yaklaşım için algoritmaların doğruluk oranları
Figure 2. *The accuracy rates of the algorithms for both approaches*

Şekil 3'te, her iki yaklaşım için algoritmaların çalışma süreleri karşılaştırılmıştır. ANN algoritması en uzun süre çalışan algoritma iken, k-NN algoritması en kısa sürede çalışan algoritma olarak bulunmuştur.

	NB	BayesNet	RF	J48	SVM	k-NN	PART	ANN	MLR
Yaklaşım I	0.03	0.05	0.37	0.06	0.06	0	0.11	2.5	0.17
Yaklaşım II	0.05	0.1	0.32	0.05	0.33	0	0.08	3.86	0.17

Şekil 3. Her iki yaklaşım için algoritmaların çalışma süreleri
Figure 3. *Running times of algorithms for both approaches*

Tüm bulgular değerlendirildiğinde, diyabet tanısının doğru sınıflandırılmasında, yüksek ayırt ediciliğe ve genel anlamda en iyi sınıflandırma kriterlerine sahip denetimli makine öğrenme algoritması RF ve BayesNet algoritmaları olmuştur.

4. Tartışma ve Sonuçlar
4. Discussion and conclusions

Özellikle sağlık alanında giderek artan hacimde veri setlerinin varlığı, makine öğrenme algoritmalarının bu alanda uygulanmasına yönelik ilgiyi arttırmaktadır. Bu nedenle sağlık alanında yapılan çalışmalarda özellikle hastalık tanısında makine öğrenme algoritmalarının kullanımı büyük önem kazanmıştır. Makine öğrenmesi yöntemleri ile oluşturulan bir tıbbi tanı modeli ile insan faktörlerinin müdahalesi hariç tutulabilmekte, geliştirilen model güçlü bir nesnelliğe sahip olabilmekte ve tıbbi tanı süreci kademeli olarak standartlaştırılarak otomatikleşebilmektedir (Chen & Pan, 2018). Bunun yanı sıra, makine öğrenmesi yöntemleri kullanılarak karmaşık veri setlerini analizinde yüksek performanslı tahminler elde edilebilmektedir (Hastie vd., 2009). Pek çok makine öğrenmesi algoritması, sadece hastalıkları sınıflandırmak veya kümelemek için değil, aynı zamanda geliştirilen modeli basitleştirmek ve modelleme sürecinde hesaplama verimliliğini artırmak için özellik seçimi için de kullanılabilmektedir (Neumann vd., 2017). Bazı araştırmacılar, diyabeti HbA1c, adiponektin ve BMI gibi klinik test verileriyle sınıflandırmak için makine öğrenme algoritmalarını kullansa da insan vücudunun karmaşıklığı nedeniyle, veri madenciliği teknolojisinin klinik tıpta uygulanması genel olarak hala nispeten sınırlıdır (Kalsch vd., 2015; Chen & Pan, 2018). Literatürde kalp, diyabet, karaciğer, kanser, kovid enfeksiyonu ve hepatit hastalıklarının tanısı için birçok farklı denetimli makine öğrenme algoritması kullanılmıştır (Fatima & Pasha, 2017; Özmen vd., 2018; Ali vd., 2021; Choudhury, 2021; Nindrea vd., 2021; Alabi vd., 2020; Muhammad vd., 2021; Tiwari, 2021). Denetimli makine öğrenmesi algoritmalarını hastalık tanısında kullanan 48 makalenin değerlendirildiği bir çalışmada, en sık uygulanan yöntemin SVM olduğu, onu NB yönteminin izlediği belirlenmiştir (Uddin vd., 2019). Fakat en yüksek doğruluk değerine RF algoritması, ikinci sırada ise SVM algoritması sağlamıştır. Benzer bir sistematik çalışmada da diyabet alanındaki makine öğrenmesi, veri

madenciliği teknikleri ve araçlarının uygulamalarının genel anlamda hangi alanlarda ve hangi algoritmalarla çalışıldığı araştırılmış ve sonuç olarak SVM, en başarılı ve yaygın olarak kullanılan algoritma olarak belirlenmiştir (Kavakiotis vd., 2017). Bizim çalışmamızda da literatürdeki çalışmalara benzer sonuçlar elde edilmiştir.

Literatürdeki çalışmalarda genellikle ortak kullanıma açık veri tabanlarındaki veri setleriyle çalışılarak, kullanılan makine öğreneme algoritmaları içerisinden en iyi performansa sahip olan algoritmanın belirlenmesi amaçlanmış ve bu çalışmalarda uzman görüşlerine yer verilmemiştir. Bununla birlikte, literatürde diyabet tanısının sınıflandırılmasında hem istatistiksel önemlilik hem de klinik önemlilik göz önünde bulundurularak modellemelerin yapıldığı bir çalışmaya rastlanmamıştır. Bu çalışmada diyabet tanısının konulmasında hangi değişkelerin daha etkili ve önemli olduğunu belirlemek için iki farklı yaklaşım temel alınarak modeller kurulmuştur. Her iki yaklaşım için diyabet tanısının doğru sınıflandırılmasında, RF ve BayesNet algoritmaları yüksek ayırt ediciliğe ve genel anlamda en iyi sınıflandırma kriterlerine sahip algoritmalardır. Algoritmaların çalışma süreleri karşılaştırıldığındaysa ANN algoritması en uzun sürede çalışan algoritma, k-NN algoritması ise en kısa sürede çalışan algoritmadır.

Sonuç olarak, bu çalışmanın, diyabet tanısının sınıflandırmasında birçok denetimli makine öğrenme algoritmalarının performanslarının değerlendirilmesinde yol gösterici olacağı düşünülmektedir. Benzer çalışmalarda modellerin performans karşılaştırılması yapılacaksa çalışmanın hassaslığını arttırması için bu tür yaklaşımlarla değerlendirilmesi önerilebilmektedir. Eğer sağlıklı bireyler, prediyabet olan bireyler ve diyabet hastası bireyler için daha eksiksiz test sonuçları kullanılırsa, diyabet tanısında glikoz veya HbA1c indeksleri anormal hale gelmeden önce tarama amaçlı bir erken tanı modeli oluşturulabilecektir. Gelecekteki çalışmalarda, diyabet hastalığının erken teşhisinde kullanılabilecek, sağlık uzmanlarının karar vermelerine destek olacak uygulamaların geliştirilmesi planlanmaktadır.

Teşekkür / Katkı Belirtme
Acknowledgement

Bu makale, 28-30 Nisan 2018 tarihlerinde Çeşme, İzmir'de (Ilıca Hotel Spa & Welness Thermal Resort) düzenlenen "4th International Researchers, Statisticians and Young Statisticians Congress (IRSYSC 2018)" kongresinde sunulan sözlü bildirinin genişletilmiş ve revize edilmiş halidir. Makalenin inceleme ve değerlendirme aşamasında yapmış oldukları katkılardan dolayı editör ve hakemlere teşekkür ederiz.

Yazar katkısı
Author contribution

Çalışma konsepti-tasarımı: Aslı SUNER, Banu SARER YÜREKLİ, Veri toplama: Banu SARER YÜREKLİ, Veri analizi ve yorumlama: Aslı SUNER, Yüksel ÖZKAN, Yazı taslağı: Aslı SUNER, Yüksel ÖZKAN, Banu SARER YÜREKLİ, İçeriğin eleştirel incelenmesi: Aslı SUNER, Yüksel ÖZKAN, Banu SARER YÜREKLİ, Son onay ve sorumluluk: Aslı SUNER, Yüksel ÖZKAN, Banu SARER YÜREKLİ

Etik beyanı
Declaration of ethical code

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. Çalışmanın etik kurul onayı, 24.12.2013 tarihinde İzmir Bozkaya Eğitim ve Araştırma Hastanesi Klinik Araştırmalar Etik Kurulu'ndan alınmış ve Helsinki bildirgesi temel alınarak çalışmaya başlanmıştır.

Çıkar Çatışması
Conflicts of interest

Yazarlar herhangi bir çıkar çatışması olmadığını beyan eder.

Kaynaklar
References

Alabi, R.O., Elmusrati, M., Sawazaki-Calone, I., Kowalski, L.P., Haglund, C., Coletta, R.D., Mäkitie, A.A., Salo, T., Almangush, A., & Leivo, I. (2020). Comparison of supervised machine learning classification techniques in prediction of locoregional recurrences in early oral tongue cancer. *International Journal of Medical Informatics, 136,* 104068. https://doi.org/10.1016/j.ijmedinf.2019.104068

Ali, M.M., Paul, B.K., Ahmed, K., Bui, F.M., Quinn, J.M.W., & Moni, M.A. (2021). Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison. *Computers in Biology and*

Medicine, *136*, 104672. https://doi.org/10.1016/j.compbiomed.2021.104672

Alpar, R. (2011). *Uygulamalı çok değişkenli istatistiksel yöntemler* (3. Baskı). Ankara: Detay.

American Diabetes Association. (2014). Standards of medical care in diabetes-2014. *Diabetes Care*, *37*, 14-80. https://doi.org/10.2337/dc14-S014

Bansal N. (2015). Prediabetes diagnosis and treatment: a review. *World Journal of Diabetes*, *6*(2), 296–303. https://doi.org/10.4239/wjd.v6.i2.296

Baratloo, A., Mostafa, H., Ahmed, N., & Gehad, E. A. (2015). Part 1: Simple definition and calculation of accuracy, sensitivity and specificity. *Emergency (Tehran, Iran)*, *3*(2), 48–49. https://doi.org/10.22037/emergency.v3i2.8154

Bilgin, G. (2021). Makine öğrenmesi algoritmaları kullanarak erken dönemde diyabet hastalığı riskinin araştırılması. *Journal of Intelligent Systems: Theory and Applications*, *4*(1), 55-64. https://doi.org/10.38016/jista.877292

Boughorbel, S., Fethi, J., & Mohammed, E. (2017). Optimal classifier for ımbalanced data using matthews correlation coefficient metric. *PloS One*, *12*(6), e0177678. https://doi.org/10.1371/journal.pone.0177678

Böhning, D. (1992). Multinomial logistic regression algorithm. *Annals of the Institude of Statistical Mathematics*, *44*(1), 197–200. https://doi.org/10.1007/BF00048682

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding rmse in the literature. *Geoscientific Model Development*, *7*, 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chen, P., & Pan, C. (2018). Diabetes classification model based on boosting algorithms. *BMC Bioinformatics*, *19*(109). https://doi.org/10.1186/s12859-018-2090-9

Choudhury A. (2021). Predicting cancer using supervised machine learning: mesothelioma. *Technology and Health Care*. *29*(1), 45-58. https://doi.org/10.3233/THC-202237

Cihan, P., & Coşkun, H. (2021). Performance comparison of machine learning models for diabetes prediction. *29th Signal Processing and Communications Applications Conference (SIU)*, (pp. 1-4). İstanbul.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Daghistani, T., & Alshammari, R. (2016). Diagnosis of diabetes by applying data mining classification techniques. *International Journal of Advanced Computer Science and Applications*, *7*(7), 329–332. https://doi.org/10.14569/IJACSA.2016.070747

Drobatz, K. J. (2009). Measures of accuracy and performance of diagnostic tests. *Journal of Veterinary Cardiology*, *11*, 33–40. https://doi.org/10.1016/j.jvc.2009.03.004

Egan, A. M., & Dinneen, S. F. (2014). What is diabetes? *Medicine (United Kingdom)*, *42*(12), 679–681. https://doi.org/10.1016/j.mpmed.2014.09.005

Er, M. B., & Işık, İ. (2021). LSTM tabanlı derin ağlar kullanılarak diyabet hastalığı tahmini. *Türk Doğa ve Fen Dergisi 10*(1), 68-74. https://doi.org/10.46810/tdfd.818528

Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, *09*(1), 1–16. https://doi.org/10.4236/jilsa.2017.91001

Goldenberg, R., & Punthakee Z. (2013). Definition, classification and diagnosis of diabetes, prediabetes and metabolic syndrome. *Canadian Journal of Diabete*s, *37*(1), 197–212. https://doi.org/10.1016/j.jcjd.2017.10.003

Goyal, A., & Mehta, R. (2012). Performance comparison of naïve bayes and j48 classification algorithms. *International Journal of Applied Engineering Research*, *7(*11 SUPPL.), 1389–1393.

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine*, *4*(2), 627–635.

Hastie, T., Tibshirani, R., & Friedman, J., (2009). *The elements of statistical learning: data mining, inference, and prediction* (Second Edition). Springer Series in Statistics.

Heikes, K. E., Eddy, D. M., Arondekar, B., & Schlessinger, L. (2008). Diabetes risk calculator: a simple tool for detecting undiagnosed diabetes and pre-diabetes. *Diabetes Care*, *31*(5),1040–1045. https://doi.org/10.2337/dc07-1150

Hripcsak, G., & Rothschild, A. S. (2005). Agreement, the f-measure, and reliability in information retrieval. *Journal of the American Medical*

Informatics Association: JAMIA, *12*(3), 296–298. https://doi.org/10.1197/jamia.M1733

International Diabetes Federation (2021). *IDF diabetes atlas*, (10th ed.) Brussels, Belgium: 2021. https://www.diabetesatlas.org.

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. https://doi.org/10.17762/ijritcc2321-8169.150361

Kaggle, 2018. (2021, November 16). http://www.kaggle.com/kumargh/pimaindiansdiabetescsv

Kalsch, J., Bechmann, L.P., Heider, D., Best, J., Manka, P., Kalsch, H., Sowa, J.P., Moebus, S., Slomiany, U., Jockel, K.H., Erbel, R., Gerken, G., & Canbay, A. (2015). Normal liver enzymes are correlated with severity of metabolic syndrome in a large population based cohort. *Scientific Reports*. *5*,13058. https://doi.org/10.1038/srep13058

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. 5th *International Journal of Computer Applications*, *45*(12), 45-50.

Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. *Computational and Structural Biotechnology Journal*, *15*, 104-116. https://doi.org/10.1016/j.csbj.2016.12.005

Li, H. Xiong, L., Ohno-Machado, L., & Jiang, X. (2014). Privacy preserving rbf kernel support vector machine. *BioMed Research International*, *2014*, 827371. https://doi.org/10.1155/2014/827371

Maniruzzaman, M., Kumar, N., Menhazul Abedin, M., Islam, M. S., Suri, H. S., El-Baz, A. S., & Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: machine learning paradigm. *Computer Methods and Programs in Biomedicin*e, *152*, 23–34. https://doi.org/10.1016/j.cmpb.2017.09.004. Epub 2017 Sep 8

Muhammad, L.J., Algehyne, E.A., Usman, S.S., Ahmad, A., Chakraborty, C., & Mohammed, I.A. (2021). Supervised machine learning models for prediction of covid-19 infection using epidemiology dataset. *SN Computer Science*, *2*(1),11. https://doi.org/10.1007/s42979-020-00394-7. Epub 2020 Nov 27

Neumann, U., Genze, N., & Heider, D. (2017). EFS: an ensemble feature selection tool implemented as R-package and web-application. *BioData Mining*, *10*, 21. https://doi.org/10.1186/s13040-017-0142-8. eCollection 2017

Nindrea, R.D., Aryandono, T., Lazuardi, L., & Dwiprahasto, I. (2018). Diagnostic accuracy of different machine learning algorithms for breast cancer risk calculation: a meta-analysis. *Asian Pacific Journal of Cancer Prevention*, *19*(7), 1747-1752. https://doi.org/10.22034/APJCP.2018.19.7.1747.

Ogurtsova, K., da Rocha Fernandes, J. D., Huang, Y., Linnenkamp, U., Guariguata, L., Cho, N. H., Cavan, D., Shaw, J. E., & Makaroff, L. E. (2017). IDF diabetes atlas: global estimates for the prevalence of diabetes for 2015 and 2040. *Diabetes Research and Clinical Practice*, *128*, 40–50. https://doi.org/10.1016/j.diabres.2017.03.024.

Olivera, A. R., Roesler, V., Iochpe, C., Schmidt, M. I., Vigo, Á., Barreto, S. M., & Duncan, B. B. (2017). Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes- Elsa-Brasil: accuracy study. *Sao Paulo Medical Journal*, *135*(3), 234–246. https://10.1590/1516-3180.2016.0309010217

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. https://doi.org/10.19113/sdufenbed.842460

Özmen, Ö., Khdr, A., & Avcı, E. (2018). Sınıflandırıcıların kalp hastalığı verileri üzerine performans karşılaştırması. *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, *30*(3), 153-159.

Tiwari, D., Bhati, B.S., Al-Turjman, F., & Nagpal, B. (2021). Pandemic coronavirus disease (Covid-19): World effects analysis and prediction using machine-learning techniques. *Expert Systems*. May 11:10.1111/exsy.12714. https://doi.org/10.1111/exsy.12714

Trevor, H., Tibshirani, R., & Friedman, J. (2011). *The elements of statistical learning: data mining, ınference and prediction* (2nd ed.). Springer International Publishing.

Uddin, S., Khan, A., Hossain, M.E., & Moni, M.A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*. *19*(1), 281. https://doi.org/10.1186/s12911-019-1004-8

Van Harmelen, F., Lifschitz, V., & Porter, B. (2008). *Handbook of knowledge representation*. Elsevier.

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. https://doi.org/10.18231/2454-9150.2018.0637

Worachartcheewan, A., Nantasenamat, C., Prasertsrithong, P., Amranan, J., Monnor, T., Chaisatit, T., Nuchpramool, W., & Prachayasittikul, V. (2013). Machine learning approaches for discerning ıntercorrelation of hematological parameters and glucose level for ıdentification of diabetes mellitus. *EXCLI Journal*, *12*, 885–893. https://doi.org/10.17877/DE290R-7572

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3

Yu, W., Liu, T., Valdez, R., Gwinn, M., & Khoury, M. J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Medical Informatics and Decision Makin*g, *10*(1), 16. https://doi.org/10.1186/1472-6947-10-16.