Makineyi Diyabetin Erken Tahmini İçin Farklı Algoritmalar Kullanarak Eğitmek ve Bunları Karşılaştırmak

Abdullah Jamous

Siber Güvenlik ve Kriptografi Bölümü, Yıldız Teknik Üniversitesi, Türkiye

Özet Diyabet, her yıl dünya çapında yüz milyonlarca insan için sağlık sorunlarına yol açmaktadır. Hastaların mevcut tıbbi kayıtları semptomları, vücut özelliklerini ve klinik laboratuvar test değerlerini ölçer; bunlar, mevcut uygulamada saptanamayan kalıpları veya özellikleri bulmayı amaçlayan biyoistatistik analizi yapmak için kullanılabilir. Bu çalışmada, diyabet hastalarının erken başlangıcını tahmin etmek için bir makine öğrenme modeli önerdik. Ayrıca, bu yöntem kullanılarak elde edilen sonuçları ve Destek Vektör Makinesi (SVM), K-En Yakın Komşu (KNN), Naïve Bayes Sınıflandırıcı (NBC) gibi birkaç geleneksel makine öğrenme algoritması yaklaşımını da karşılaştırdık. Önerdiğimiz yöntemin hesaplama sonuçları, çok daha fazla özelliğe ihtiyaç duyulduğunu ve sahip olduğumuzdan daha fazla örnek elde edildiğinde daha yüksek tahmin doğruluğunun elde edilebileceğini göstermektedir. En yüksek başarı oranını %84 ile bu deneyde elde ettik. Bu çalışmanın klinik pratiğe uygulanabilme potansiyeline sahip olması ve hekim/doktorlar için destekleyici bir araç haline gelmesi için biraz geliştirilmesi gerekmektedir.

Anahtar Kelimeler makine öğrenmesi, kümeleme, başarı oranı, modül.

I. Giriş

Diyabet kronik ve dünya çapında bir hastalıktır. 2012 yılında gelişmekte olan ülkelerde ve özellikle gelişmiş ülkelerde önde gelen sekiz ölüm nedeninden biridir. Bu nedenle hükümetler ve bireyler bütçelerinin büyük bir bölümünü bu ciddi hastalığı araştırmak ve tedavi bulmak için harcamaktadır [1]. Diyabet, kan şekeri metabolizmasında bir bozukluğa yol açan bir eksiklik veya insülin direnci nedeniyle kan şekerinizin sürekli olarak normalden

yüksek olduğu bir durumdur. Hastalar şeker hastası olduğunda, vücutları tükettikleri çoğu karbonhidratı parçalayamaz veya günlük aktiviteler için enerji üretmek üzere verimli bir şekilde şeker glikozuna dönüştüremez. Bu nedenle, kan dolaşımında kademeli olarak şeker birikmesine neden olur. Glikoz daha sonra kan dolaşımında kalır ve vücuttaki her hücreye ulaşmaz. İki tür hastalık vardır: tip 1 ve tip 2. Tip 1 diyabet, doğumda doğuştan gelen bir genital kusur nedeniyle pankreasın yeterince insülin üretmemesi veya

üretememesi nedeniyle insülin eksikliğinin ortaya çıkmasıdır. Tip 1 diyabet nadirdir, genellikle küçük çocuklarda görülür. Tip 2 ile hastalar insüline dirençlidir. Bu, vücudun hala üretebileceği, ancak insülin glikozu metabolize edemediği anlamına gelir. Dünyadaki diyabet hastalarının yaklaşık %90'ı 2'dir [2]. Uluslararası Divabet Federasyonu'na (IDF) [3] göre, 20 ila 79 yaşları arasında yaklaşık 463 milyon yetişkin diyabetle yaşıyor, 2045 yılına kadar diyabetli insan sayısı yaklaşık 700 milyona çıkıyor. 2019 yılında diyabet en az 760 milyar dolara neden oldu - yetişkinler toplam harcamanın %10'unu olusturuyor. yaklasık Divabet hastalarını erken bir aşamada tahmin etmek için farklı parametreleri birleştirmek gerektiğinden, diyabet teşhisi zor bir problem olarak kabul edilir. Günümüzde, divabet hastası olan hastaların sayısı banliyölerde giderek artmakta ve kentsel alanlarda dramatik bir şekilde artmaktadır. Bu nedenle ne kadar erken teshis konulursa hastalar için o kadar iyi Bununla birlikte. olacaktır. sürecin karmaşıklığı ve diğer semptomlar nedeniyle hastalarda olabilir. Diyabetin erken teşhisinin zorluğu nedeniyle tedavi operasyonunun ertelenmesine neden olabilir. Bu nedenle, tip alanında çalışan herkesin daha hızlı ve doğru koşullara sahip hastalara teşhis koymasını desteklemek için erken bir diyabet tahmin sistemi geliştirmek çok önemlidir. uzmanlarının diyabet veya kalp yetmezliği gibi çeşitli hastalıkları teşhis etmesine yardımcı olmak için makine öğrenimi ve derin öğrenme algoritmaları başarıyla uygulandı. ANN tıp araştırmacılar tarafından alanındaki uygulanmıştır [4], [5]. Sürekli gelişen teknoloji çağında bilgisayar teknolojisi, hastalıkların erken ve doğru bir şekilde teşhis edilmesine yardımcı olarak zamandan ve paradan tasarruf sağlayabilir. Birçok araştırmacı hastalığı teşhis etmek için veri madenciliği yöntemlerini uygulamıştır. K-En Yakın Komşu Sınıflandırıcı (KNN) [6], Naive Bayes

Sınıflandırıcı (NBC) [7] gibi diğer standart makine öğrenimi/derin öğrenme algoritmaları karşılaştırmadır. Destek Vektör Makinesi (SVM) [8], Karar Ağacı (DT) [9] ve Rastgele Orman Sınıflandırıcısı (RFC) [10]. [11]'ye göre Tao Zheng, Tip 2'yi keşfetmek için KNN, DT, NBC, Rastgele Orman sınıflandırıcısı (RFC), Lojistik Regresyon (LR), Destek Vektör Makinesi (SVM) dahil olmak üzere bazı makine öğrenimi modellerini önerdi. Diabetes mellitus (T2DM) Yazar 300 hasta örneğinden toplamıştır. LR ile %99, NB ile %96, RF ile %98, KNN ile %97, SVM ile %96, J48 ile %98 doğruluk buldular. [12]'teki vazarlar, UCI deposunda bulunan Pima Indians Diabetes veri setini kullanmıslardır. Bu veri kümesi 768 örnek ve sekiz öznitelik içerir. Diyabet hastalarını tespit etmek için DT, SVM. NBC gibi üç makine öğrenimi sınıflandırması uyguladılar. NBC, modellere kıyasla en yüksek doğruluğu %76,30 ile elde etmektedir. Diğer bir yöntem ise Makine Öğrenimi Torbalama Topluluğu Sınıflandırıcısıdır (ML-BEC) [13]. Arastırmacının diyabeti teshis etmek ve erken bir aşamada tahmin etmek için bu yöntemi kullanması zorunludur. ML-BEC'in asamasında, t-dağılımlı Stokastik tekniği uygulanarak özellik çıkarımı da yapılmaktadır. Öte yandan, [14]'teki yazar Pima Indians Diabetes veri setini (738 hasta) kullanmıştır. Diyabet hastalarını teşhis etmek için yazarlar, bu veri setindeki performansı değerlendirmek için SVM, KNN, NBC, ID3, C4.5, CART gibi modelleri uyguladılar. SVM ve LDA algoritmaları %88 ile en iyi doğruluğu elde Aksine Rabina divabet etmistir. [15], hastalarını teşhis etmek için denetimli ve denetimsiz öğrenmeyi uygulamıştır. En iyi sınıflandırma modellerini bulmak için WEKA kullandılar ve Karar aracını Ağacı diğerlerinden daha iyi performans gösterdi. [16]'de Tejas N. Joshi, diyabet hastalarının daha erken saptanması için üç teknik önermiştir. Yazar, SVM, Lojistik regresyon ve

ANN dahil olmak üzere makine öğrenimi modellerini veri setindeki 7 öznitelikle (Glikoz, Kan Basıncı, Cilt Kalınlığı, İnsülin, BMI, Diyabet Soy İşlevi ve yaş) karşılaştırdı. Bu modeller uygulandıktan sonra Destek Vektör Makinesinin (SVM) performansı diğer sınıflandırma yöntemlerine göre daha iyidir.



ŞEKİL 1. Filtre yöntemi süreci.

II. Sistem Tasarımı: A. VERİ ÖN İSLEME

Bu veri seti, bir hastanede doğrudan ve hastaların bilgilerinden manuel olarak hazırlanmıştır. Tüm bu veriler, Irak'ın Erbil sehrinde bulunan özel bir hastane ve kalp merkezi olan Medical Help Center'da toplanmıştır. Bu veri setini oluşturmanın temel amacı, bir hastanın kalp hastalığından muzdarip olup olmadığını tahmin etmek için yerli hastaların kalp sorunları hakkındaki bilgileri üzerinde çalışmaktır. İlk olarak, doktora gelen hastalar için veriler asistan doktorlar ve kardiyologlar tarafından manuel olarak kağıt sayfalara yazılmıştır. Daha sonra tüm veriler araştırmacı tarafından çok dikkatli ve doğru bir şekilde excel dosyasına kaydedilir. Toplanan veriler beş kategoriye ayrılmıştır. Bazıları hastanın demografik bilgileri, bazıları hastanın öyküsü, bazıları fizik muayene ve semptomatik, bazıları tıbbi laboratuvar testleri ve bazıları da tanısal özelliklerdir.

Veri setinin özellikleri tıp doktorlarının tavsiyelerinden seçilmiştir. Kalp hastalığı ile ilgili etkili özellikler için araştırma makalelerinin önerileri de dikkate alınır. Kardiyovasküler hastalığa katkıda bulunan ortak faktörler veya özellikler tanımlanmıştır. "Hedef" alanı, hastada kalp hastalığının varlığına atıfta bulunur. Düzgün bir şekilde tamamlandıktan sonra hastaların isimleri veri

tabanından kaldırıldı.

B. MAKİNE ÖĞRENME MODELLERİ

1) K - NEAREST NEIGHBOR (KNN)

KNN, Veri Madenciliği alanında çok rağbet görmektedir ve Eğitim Verileri'nde (Sorgu noktası) tüm nesneler ile nesneler arasındaki en yakın mesafeye göre nesneleri sınıflandırmak için kullanılmaktadır [6]. Bir öğe, K komşusuna göre sınıflandırılır. K, algoritmayı gerçekleştirmeden önce belirlenen pozitif bir tam sayıdır. İnsanlar genellikle nesneler arasındaki mesafeyi hesaplamak için Öklid mesafesini kullanır. Öklid mesafesi aşağıdaki denklemde hesaplanabilir:

$$d_{Euclidean}(x, y) = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

KNN yöntemi aşağıda açıklanmıştır:

- K değerini belirleyin (en yakın komşu sayısı)
- Sınıflandırılmış nesne (Sorgu Noktası) ile tüm nesneler arasındaki mesafeyi hesaplayın.
- Boşlukları artan sırada düzenleyin ve Sorgu Noktasına en yakın K komşuyu bulun.
- Tanımlanmış en yakın K komşunun tüm sınıflarını alın.
- Sorgu Noktası için sınıfı belirlemek için en yakın komşunun sınıfının çoğuna güvenmek.

2) DESTEK VEKTÖR MAKİNASI (SVM) SVM, denetimli makine öğrenimi modelinin bir biçimidir [8]. Daha az aykırı değer içeren nispeten küçük bir veri seti için uygundur. Fikir, veri noktalarını ayırmak için bir hiper şerit bulmaktır. Bu hiper düzlem, alanı farklı alanlara bölecek ve her alan bir tür veri içerecektir. İki veri sınıfını ayırmak için seçilebilecek birçok hiper düzlem vardır. Hedefimiz maksimum marja sahip bir uçak bulmaktır. Kenar bosluğu, hiperdüzlem ile iki

alt sınıfa karşılık gelen en yakın iki veri

noktası arasındaki mesafedir. DVM, bu marj değerini maksimize ederek algoritma ritmini optimize etmeye çalışır, böylece verileri iki katmana bölmek için en iyi süper düzlemi bulur. Hiper düzleme en yakın olan veri noktalarına destek vektörleri denir. Hiperdüzlem, uzayı iki bölüme ayıran doğrusal bir yüzeydir. Bir hiperdüzlem, bir ikili sınıflandırıcıdır ve ortam uzayından bir boyutu daha az olan bir alt uzaydır.

3) NAÏVE BAYESIAN CLASSIFIER (NBC)

NBC, gözlemlenen ve istatistiksel verilere dayalı olarak yargılarda bulunmak ve verileri sınıflandırmak için olasılık teorisinin Bayes teoremine dayanan bir algoritmadır [7]. NBC, anlaşılması görece kolay ve oldukça doğru olduğundan, toplanan bir veri kümesine dayalı olarak en doğru tahminleri yapmak için sıklıkla kullanılan algoritma ritimlerinden biridir. Denetimli Makine Öğrenimi Algoritmaları grubuna aittir.

Bayes teoremine göre, x değerine bağlı olarak y değerinin rastgele olasılığını hesaplamak için kullanılan denklem:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

x ile, x = (x1,x2,x3,...,xn) şeklinde yazılabilen bir vektördür. O zaman Bayes denklemi söyle olur:

$$P(y|x1...,xn) = \frac{P(x_1|y) P(x_2|y)...P(x_n|y)P(y)}{P(x_1) P(x_2)...P(x_n)}$$

Ardından, P (y \mid X)'in maksimuma ulaşması için hedef sonuç y şöyle olur:

$$y = argmax_y P(y) \prod_{i=1}^{n} p(x_i|y)$$

4) Gaussian Naive Bayes (GaussianNB)

Sürekli değerleri işlemek için başka bir yaygın teknik, Bernoulli tarafından dağıtılan yeni bir özellik seti elde etmek için özellik değerlerini ayrık hale getirmek için binning kullanmaktır; Aslında bazı literatür, saf Bayes'i uygulamak

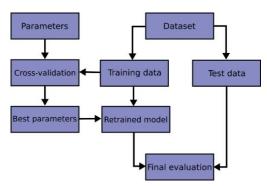
için bunun gerekli olduğunu öne sürüyor, ancak öyle değil ve ayrıklaştırma, ayrımcı bilgileri çöpe atabilir.

Bazen sınıf-koşullu marjinal yoğunlukların dağılımı normal olmaktan uzaktır. Bu durumlarda, çekirdek yoğunluğu tahmini, her sınıfın marjinal yoğunluklarının daha gerçekçi bir tahmini için kullanılabilir. John ve Langley tarafından tanıtılan bu yöntem, sınıflandırıcının doğruluğunu önemli ölçüde artırabilir.

$$p(x=v\mid C_k) = rac{1}{\sqrt{2\pi\sigma_k^2}}\,e^{-rac{(v-\mu_k)^2}{2\sigma_k^2}}$$

III. Deneysel Analiz

Verileri alıp içeriğine baktıktan sonra boş değerler içeren 18 satır bulundu.Bu değerleri içeren sütunun önemi (yaş özelliği) göz önüne alındığında, tüm bu satırları silmeye karar verdim.Daha sonra hesapladım aykırı değeri her sütun için Bu işlem, yüksek dağılım oranına sahip Sütunları bulmamıza yardımcı olur.



SEKİL 2. Veri analiz süreci.

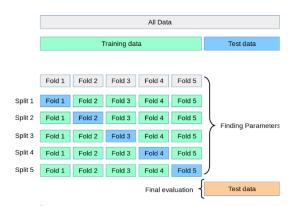
Aşağıdaki şekilde gösterildiği gibi dağılım değeri verilerine bakıldığında:

hr: 0.7461234232795414 0.10955774412687547 sex: htn: 0.053694655223828276 0.028891474200243034 chp: bpdias: 0.025738421944040895 ivsd: 0.008029896516294224 0.008021157414321833 bpsys: smoke: 0.004864994617797069 0.00434391552477463 ld1: 0.00210627235648458 years: fh: 0.0020648438807396967 0.0018241179903409161 weight: 0.0013291344402447118 height: 0.0013039661008417275 lifestyle: 0.0006236208480725152 dm: active: 0.0004603499284436999 0.0004585573261552263 ihd: 0.00039894995254545165 age: 0.00010996051049576378 awave: ecgpatt: 5.45438179189075e-05

Bu nedenle dağılım değeri 0.0001'den küçük olan (ecgpatt) özelliği olan sütunu silmeye karar verdim.

Verileri sınıflandırmak için eğitim grubu ve deney grubu olarak ayırdım ve dağılım şu şekildeydi: test grubu için %15, eğitim grubu için %85. Daha önce de belirttiğim gibi (GaussianNB), (SVC), (KNN), (NBC) olmak üzere 4 algoritma, daha sonra her işlem için başarı oranını buldum.

Daha sonra eğitim grubunu birkaç gruba ayırdım ve en yüksek başarı oranını bulmak için Cross-Validation denilen kullandım. Bu işlemi 8 kez tekrar test ettim ve her seferinde parça sayısını artırdım. her model: SVC için 4 parça, KNN için 9 parça, GaussianNB için 4 parça ve NBC için 5 parça.



SEKIL 3. Cross-Validation.

Bir tahmin fonksiyonunun parametrelerini öğrenmek ve bunu aynı veriler üzerinde test etmek metodolojik bir hatadır: az önce gördüğü örneklerin etiketlerini tekrar eden bir model mükemmel bir skora sahip olur ancak henüz yararlı bir şey tahmin edemez. görülmemiş veriler Bu duruma aşırı uyum denir. Bundan kaçınmak için, (denetimli) bir makine öğrenimi deneyi gerçekleştirirken mevcut verilerin bir kısmını bir test seti X test, y test olarak tutmak yaygın bir uygulamadır. kelimesinin yalnızca akademik "Deney" kullanım için tasarlanmadığını unutmayın, çünkü ticari ortamlarda bile makine öğrenimi genellikle deneysel olarak başlar. Model eğitiminde tipik çapraz doğrulama iş akışının akış şemasını burada bulabilirsiniz. En ivi parametreler grid search teknikleri belirlenebilir.

IV. Sonuç

Bu bölüm, özel özellik seçme yöntemlerinin uygulanmasından elde edilen denevsel sonuçları ve diğer algoritmalarla karşılaştırmayı özetlemektedir. Sonraki üç alt bölüm, sistem performansının sonuçlarını ve değerlendirmesini içerir. Bu sistemde veri seti 85:15'ye bölünerek verilerin %85'i modelleri eğitmek, %15'si ise bu modellerin doğruluğunu test etmek için kullanılmaktadır. Bu algoritmaların performansı, aşağıdaki denklemlerde verilen Doğruluk, Doğruluk, Geri Alınabilirlik, F1 puanı gibi performans metriklerine dayalı olarak incelenir:

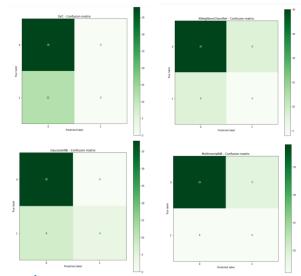
Accuracy =
$$\frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F_1 = 2*\frac{Precision * Recall}{Precision + Recall}$$

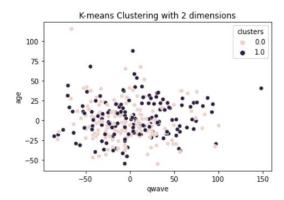
- Gerçek Pozitif (TP): numuneler (T) iken doğru (T) olarak sınıflandırılır.
- Gerçek Negatif (TN): numuneler (F) iken yanlış (F) olarak sınıflandırılır.
- Yanlış Pozitif (FP): numuneler (F) iken (T) olarak sınıflandırılır.
- Yanlış Negatif (FN)



SEKİL 4. Confusion matris.

Bu çalışmada KNN, SVM, NB, GaussianNB gibi 4 sınıflandırıcı modeli geliştirilmiştir. Bu verileri eğitmeden önce aykırı veri kümesinin uyguluyoruz. kaldırılmasını Asağıdaki şekildeki çubuk grafik, makine öğrenimi algoritmaları arasındaki doğruluk karşılaştırmasının göstergesini göstermektedir. Şekil 4'ten de görülebileceği gibi KNN ve RFC %84 ile en yüksek doğruluğa sahiptir. SVM ve DT de diğerlerine kıyasla iyi bir doğruluk elde eder. Matrisin köşegen elemanları, her veri katmanı için doğru şekilde sınıflandırılmış nokta sayısıdır.

Doğruluk, köşegen üzerindeki bu bileşenlerin toplamının tüm matrisin öğelerinin toplamına bölünmesiyle çıkarılabilir. İyi bir model, faktörlerle bir karışıklık matrisi verecektir.



SEKİL 4. K-Means Kümeleme.

Bu bölümde, sonuçlar %40'ı geçmeyen düşük bir başarı oranı gösterdiğinden ve bu yüzde bu tür algoritmalar için çok düşük olduğundan ve düşüşün nedeni muhtemelen olduğundan, algoritma grup kavramını uygulamayı basaramamıstır. deneyler Bu tür için verilerdeki satır sayısı çok az olduğundan, daha yüksek bir başarı oranı elde etmeye yardımcı olmak için bundan daha fazla veriye sahip olmanızı ve daha spesifik özellikler eklemenizi öneririm.

REFERANSLAR

- [1] D. Falvo and B. E. Holland, Medical and Psychosocial Aspects of Chronic Illness and Disability. Burlington, MA, USA: Jones & Bartlett Learning, 2017.
- [2] G. Klöppel, M. Löhr, K. Habich, M. Oberholzer, and P. U. Heitz, "Islet pathology and the pathogenesis of type 1 and type 2 diabetes mellitus revisited," Pathol. Immunopathology Res., vol. 4, no. 2, pp. 110–125, 1985.
- [3] International Diabetes Federation—Facts & Figures. Available: https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html.
- [4] C. S. Dangare and S. S. Apte, "A data mining approach for prediction of heart disease using neural networks," ResearchGate, vol. 3, no. 3, pp. 30–40, 2012.

- [5] S. Smiley. (Jan. 12, 2020). Diagnostic for Heart Disease with Machine Learning. Medium. Available: https://towardsdatascience.com/diagnostic-for-heart-disease-with-machine learning-81b064a3c1dd.
- [6] An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression: The American Statistician. Available: https://www.tandfonline.com/doi/abs/10.1080/00031305.
- [7] K. M. Ting and Z. Zheng, "Improving the performance of boosting for naive Bayesian classification," in Proc. Methodol. Knowl. Discovery Data Mining, Berlin, Germany, 1999, pp. 296–305.
- [8] N. V. Vapnik, Statistical Learning Theory. Hoboken, NJ, USA: Wiley, Sep. 1998. Accessed Sep. 6, 2020.
- [9] J. R. Quinlan, "Induction of decision trees," Mach. Learn., vol. 1, no. 1, pp. 81–106, Mar. 1986.
- [10] L. Breiman, "Random forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, Oct. 2001.
- [11] T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, and Y. Chen, "A machine learning-based framework to identify type 2 diabetes through electronic health records," Int. J. Med. Informat., vol. 97, pp. 120–127, Jan. 2017.
- [12] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," Procedia Comput. Sci., vol. 132, pp. 1578–1585, Jan. 2018.
- [13] S. K. Somasundaram and P. Alli, "A machine learning ensemble classifier for early prediction of diabetic retinopathy," J. Med. Syst., vol. 41, no. 12, p. 201, Nov. 2017.
- [14] P. Agrawal and A. k. Dewangan. (2015). A Brief Survey On The Techniques Used For The Diagnosis Of Diabetes-Mellitus. Available: paper/A-BRIEF-SURVEY-ON-THE-TECHNIQUES-USED-FOR-THE-OF-Agrawal-Dewangan/198dd36f4a84386817693eaa55b600005e059abd
- [15] E. Rabina and E. A. Chopra. (2016). Diabetes Prediction by Supervised and Unsupervised Learning With Feature Selection. Available: paper/Diabetes-Prediction-by-Supervised-and-Unsupervised-Rabina-Chopra/ 105093e74f7720ac14ddafa1790ca9a48f119d6e
- [16] T. N. Joshi and P. N. Chawan, "Diabetes prediction using machine learning techniques," Int. J. Eng. Res. Appl., vol. 8, no. 1.



Abdullah Jamous lisans derecesini 2022 yılında Sakarya Üniversitesi Bilgisayar Mühendisliği Fakültesi'nden almıştır. Şu anda Yıldız Teknik Üniversitesi'nde Siber Güvenlik ve Kriptografi alanında yüksek lisans programına devam etmekte ve aynı zamanda tam yığın programcıdır. İlgi alanları Gömülü Sistemler, Nesnelerin İnterneti, Yapay Zeka Uygulamaları ve Kriptografi Uygulamalarıdır.