TEKNOFEST

HAVACILIK, UZAY VE TEKNOLOJİ FESTİVALİ

SAĞLIKTA YAPAY ZEKÂ YARIŞMASI

(Medikal Teknolojiler Kategorisi)

PROJE SUNUŞ RAPORU

TAKIM ADI

Aspect-AI

BAŞVURU ID

326022

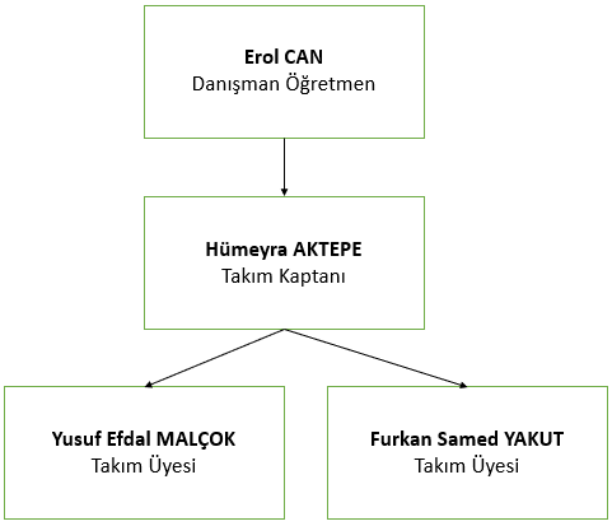
İçindekiler

Takım Şeması **3**
Proje Özeti **3**
Proje Tanımı ve Özgün Değer **4**
Yöntem **5**
A. Sınıflandırma Algoritmaları 5
1. Lojistik Regresyon 5
2. Karar Ağaçları 6
3. K-En Yakın Komşu 6
B. Veri Önişleme 6
C. Performans Metrikleri 7
D. Uygulama Geliştirme 8
E. Deneysel Sonuçlar 8
Riskler ve B Planları **8**
Referanslar **9**

Proje Konusu

☒ Hastalık teşhisine yönelik karar destek sistemleri

1. Takım Şeması



Aspect-AI Takımında 1 danışman öğretmen ve 3 öğrenci yer almaktadır. Öğrencilerden biri takım kaptanı olarak proje ile ilgili iş ve işlemlerden sorumludur. Takım kaptanı ile birlikte diğer iki öğrenci ise veri setini tanıma ve veri ön işleme adımlarından, yapay zeka algoritmalarının uygulanmasından ve sonuçların analizinden sorumludur.

2. Proje Özeti

Projenin amacı; Yapay zeka yöntemleri kullanılarak makine öğrenmesi ile diyabet hastalığının tahmini için bir model oluşturmak ve kullanılan algoritmaları karşılaştırmak.

Projenin konusuna yönelik çözüm önerileri; Makine öğrenmesi neticesinde oluşan model ile ilgili veriler girilerek diyabet hastalığını tahmin eden model, hekimlerimize hızlı bir şekilde hastalıkla ilgili dönüş yaparak erken teşhis ile can kayıplarının önüne geçmek ve insanların sağlığına kavuşmasına yardımcı olmak istiyoruz. Bu nedenle en uygun olan yapay zeka algoritmasının hangisi olduğunu analiz ederek modelimizin başarım oranına göre uygun algoritmayı seçmek istedik.

Özgün Yanları; Projemizde 3 adet sınıflandırma algoritması çalıştırılmış ve karışıklık matrisi ile en uygun algoritmanın hangisi olduğu gözlemlenmiştir.

Kullanılan Yöntemler; Projemizde, sınıflandırma algoritmalarından Lojistik Regresyon, Karar Ağaçları ve K-En Yakın Komşu algoritmaları çalıştırılmıştır. Bu üç algoritma içinden %63 f-score ve %77 doğruluk payı ile K-En Yakın algoritması daha iyi performans sergilemiştir.

3. Problem Tanımı ve Özgün Değer

Diyabet, dünyada giderek artan bir sağlık sorunu haline gelmiştir. Her hastalıkta olduğu gibi diyabet hastalığında da erken teşhis oldukça önemli bir etkendir. Erken teşhis ile hastalığın tedavisi ve iyileşme süreci hızlanmaktadır. Bu sorunu anlamak ve hedefe yönelik çözümler tasarlamak için daha fazla araştırmaya ihtiyaç duyulmuştur. [1] Yapay zeka alanındaki gelişmeler ile bu sorunun çözümüne katkıda bulunmak için diyabet veri seti üzerinde sınıflandırma algoritmalarını çalıştırmış ve karışıklık matrisi ile en iyi performansa sahip algoritmayı belirleyerek bir model oluşturmayı başarmış bulunmaktayız. Modele gönderilen yeni veri nesnesi hakkında yorum yapabilme ve tahminde bulunma sağlanmaktadır. Literatürde diyabet hastalığının teşhisi üzerine pek çok çalışma bulunmaktadır.

Çoşkun ve Baykal SEER veri kümesi üzerinden J48 karar ağacını, NB, LR ve k-star algoritmalarını karşılaştırmışlardır. Deneyler sonucunda J48 algoritmasının %86.36 doğruluk payı ile en iyi sonucu verdiğini gözlemlemişlerdir [2]. Başer, Yangın ve Sarıdaş 130 hastanedeki 70000 hastaya ait veriler üzerine NB, knn, LR, karar ağaçları ve rastgele orman uygulamışlar ve %84.78 ile bir topluluk öğrenmesi algoritması olan rastgele ormanın en başarılı olduğu sonucuna varmışlardır [3]. Bir diğer çalışmada Pima Indian Diyabet veri kümesi kullanılmıştır. Bu çalışma için seçilen veri madenciliği yöntemlerinden knn, rastgele orman, DVM, yapay sinir ağları (YSA) ve karar ağaçlarını kullanmışlar ve bu çalışmada da rastgele orman daha yüksek doğruluk ile sınıflandırma sağlamıştır [4]. Bir diğer çalışmada PIDD veri kümesi üzerinde gradient bossting, LR ve NB algoritmaları kullanılmış ve gradient goosting algoritmasının %86 oranıyla daha başarılı sonuç verdiği tespit edilmiştir [5]. Yine PIDD veri kümesi üzerine yapılan bir diğer çalışmada LR, karar ağaçları, rastgele orman, DVM ve AdaBoost algoritmaları karşılaştırılmış ve en iyi performansa sahip algoritmanın %87.1 ile LR olduğunu gözlemlenmiştir [6]. Day, Hossain ve Rahman 2019 yılında, DVM, knn, nb ve YSA ile yapılan deneylerde en başarılı algoritmanın YSA algoritması olduğunu görmüşlerdir [7]. Murthy ve Srilatha karar ağaçları ve LR algoritmalarını karşılaştırmış olup; %82.46 başarı oranı ile LR algoritmasının başarılı sonuç verdiği tespit etmişlerdir [8].

4. Yöntem

Deneysel çalışmada kullanılan veri kümesi 768 örnekten ve 8 öznitelikten oluşmaktadır. Veri kümesini oluşturan öznitelikler ve aldıkları değerler Tablo I'de verilmiştir. Tüm öznitelikler gerçek sayı değeri almaktadır. Veri kümesi %70'i eğitim kümesi, %30'u test kümesi olarak ayrılmıştır. Eğitim kümesinde 537 adet örnek bulunurken, test kümesinde 231 adet örnek bulunmaktadır. Ayrıca eldeki veri kümesi sınıf dağılımı açısından dengesizlik göstermektedir. Diyabet olmayan hasta sayısı 500 iken diyabetli hasta sayısı 268'dir.

TABLO I. VERİ KÜMESİ

Öznitelik	Tanımı	Değer Aralığı
Pregnancies	Bir kişinin hamile kalma sayısını göstermektedir	0-17
Glucose	Oral glukoz tolerans testi'den 2 saat sonra ölçülen kan glukozu konsantrasyonları	0-199
Bloodpressure (mm Hg)	Diyastolik kan basıncı	0-122
SkinThickness (mm)	Triceps deri kıvrım kalınlığı	0-99
Insulin (mu U/ml)	2 saatlik serum insülini	0-846
BMI (kg/m^2)	Vücut kitle endeksi	0-67.1
DiabetesPedigreeFunction	Diyabet soyağacı işlevi	0.08-2.42
Age (Yıl)	Yaş	21-81
Outcome	0 kişi diyabet değil, 1 kişi diyabet	0,1

A. *Sınıflandırma Algoritmaları*

1) *Lojistik Regresyon*

Lojistik Regresyon(LR), bir veya daha fazla tahmin değişkenine dayalı olarak kategorik bir değişkeni tahmin etmek için regresyon analizini kullanan sınıflandırma yöntemidir. Lojistik Regresyon bağımlı değişkenin değerinin bağımsız değişkenler kullanılarak tahmin edildiği varsayımına dayanmaktadır.

2) Karar Ağaçları

Karar ağaçlarının ilk hücrelerine **kök** (root veya root node) denir. Her bir gözlem kökteki koşula göre "Evet" veya "Hayır" olarak sınıflandırılır. Kök hücrelerinin altında **düğümler** (interval nodes veya nodes) bulunur. Her bir gözlem düğümler yardımıyla sınıflandırılır. Düğüm sayısı arttıkça modelin karmaşıklığı da artar. Karar ağacının en altında **yapraklar** (leaf nodes veya leaves) bulunur. Yapraklar, bize sonucu verir.

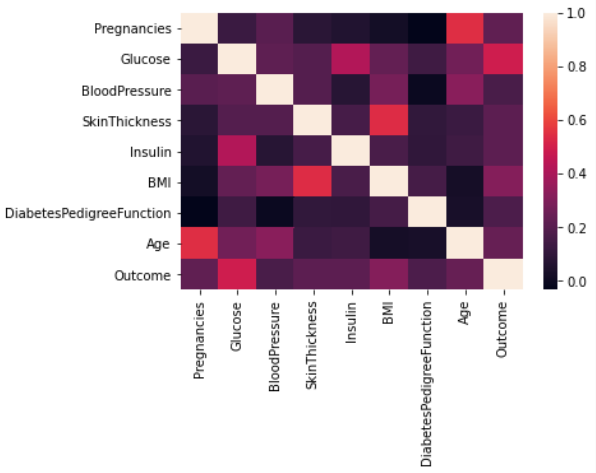
3) K-En Yakın Komşu

K-En Yakın Komşu Algoritması; eski, basit ve gürültülü eğitim verilerine karşı dirençli olması sebebiyle en popüler makine öğrenme algoritmalarından biridir. Fakat bunun yanında dezavantajı da mevcuttur. Örneğin, uzaklık hesabı yaparken bütün durumları sakladığından, büyük veriler için kullanıldığında çok sayıda bellek alanına gereksinim duymaktadır.

B. Veri Önişleme

Veri kümesine, veri önişleme aşamasında ilk olarak boş kayıtların olup olmadığı gözlenmiştir ve herhangi bir boş kayda rastlanmamıştır. İkinci aşama gürültülü verilerin olup olmadığı araştırılmıştır ve beş adet sıfır değerinin olduğu görülmüştür. Glucose özniteliğinin sıfır değeri alamayacağı göz önünde bulundurulduğunda bu değerlerin gürültülü olduğu anlaşılmıştır. Bu durumdaki kayıtlar için sıfır değeri Glucose özniteliğinin aldığı değerlerin ortalaması ile değiştirilmiştir.

Önişleme adımlarından bir diğeri ise veri kümesinde boyut azaltma olarak gerçekleştirilmiştir. Buradaki amaç veri kümesini daha az sayıda öznitelik ile temsil ederek hem bir hastalığın teşhisinde asıl rol oynayan öznitelikleri belirlemek böylece tanı süresini azaltmak hem de sınıflandırma modellerinin başarısını arttırmaktır. Bu amaçla korelasyon katsayısından faydalanılmıştır. Öznitelikler arasındaki ilişkiyi gösteren korelasyon tablosu Şekil 1 ile gösterilmiştir.



Şekil 1. Öznitelikler ararsındaki korelasyon ilişkisi.

Korelasyon tablosu incelendiğinde 8 adet öznitelikten Glucose, BMI ve Age olmak üzere 3 tanesi seçilmiştir.

C. Performans Metrikleri

Sınıflandırma modelinin başarımını değerlendirmek için kullanılan temel ölçütler doğruluk, kesinlik, duyarlılık ve F-score ölçütleridir. Eldeki veri kümesi üzerinden performans metriklerini tanımlayacak olursak; doğruluk doğru sınıflandırılan örneklerin (diyabet/diyabet olmayan) tüm örneklere oranıdır. Kesinlik diyabet olan ve doğru sınıflandırılan örneklerin diyabet olarak sınıflandırılan örneklere oranıdır. Duyarlılık diyabet olan ve doğru sınıflandırılan örneklerin tüm diyabet olanlara oranıdır. F-Score, kesinlik ve duyarlılığın harmonik ortalamasıdır ve ikisinin iyi bir kombinasyonunu vermeye çalışır. Bu ölçütleri hesaplayabilmek için Tablo II'deki karışıklık matrisinden faydalanılır.

TABLO II. KARIŞIKLIK MATRİSİ

		Gerçek Sınıf	
		Sınıf=1	Sınıf=0
Tahmin Edilen Sınıf	Sınıf=1	TP	FP
	Sınıf=0	FN	TN

Tablo II üzerinden performans metriklerinin formülleri aşağıdaki eşitlikler ile verilmiştir.

$$Doğruluk = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$Kesinlik = \frac{TP}{TP + FP} \tag{2}$$

$$Duyarlılık = \frac{TP}{TP + FN} \tag{3}$$

$$F - Score = \frac{2 \times Kesinlik \times Duyarlılık}{Kesinlik + Duyarlılık} \tag{4}$$

D. Uygulama Geliştirme

Geliştirilen yapay zeka modellerinden en iyi başarım oranına sahip olan K-En Yakın Komşu algoritması kullanılarak Python programlama dili ile bir API yazılmış ve bu API'ye de geliştirilen mobil uygulama bağlanmıştır. Mobil Uygulama üzerinden gerekli veriler girildiğinde API altında çalışan model, kişinin diyabet hastası olup olmadığı konusunda geri bildirim sağlamaktadır.

E. Deneysel Sonuçlar

Algoritmaların çalıştırılması sonucunda elde edilen metrik değerleri aşağıda gösterilmiştir.

TABLO III. ALGORİTMALARIN METRİK DEĞERLERİ

Algoritma Adı	**Metrik Değerleri**
Lojistik Regresyon (LR)	[[116 14] [29 33]]
Karar Ağaçları (KA)	[[106 24] [24 38]]
K-En Yakın Komşu (KNN)	[[112 18] [25 37]]

Uygulanan Sınıflandırma algoritmalarına göre F-Score sonuçları Lojistik Regresyon 0.60, Karar Ağaçları 0.61 ve K-En Yakın Komşu 0.63 olarak gözlemlenmiştir. Doğruluk katsayıları ise Lojistik Regresyon 0.77, Karar Ağacı 0.75 ve K-En Yakın Komşu ise 0.77 olarak gözlemlenmiştir.

5. Riskler ve B Planları

Kullanılan veri seti, çalıştırılan algoritmalar, eğitim ve test kümelerinin oransal paylaşımları ve uygulayıcı farklılığı gibi etkenler algoritmaların başarım sonuçlarını etkilemektedir. Dolayısıyla her bir uygulama için farklı başarım oranlarıyla karşılaşmak olağan bir durumdur.

6. Referanslar

[1] S. I. Omurca and E. Ekinci, “An alternative evaluation of post traumatic stress disorder with machine learning methods,” Sep. 2015, doi: 10.1109/INISTA.2015.7276754.

[2] C. Coşkun, A. B.-A. Bilişim, and U. 2011, “Veri madenciliğinde sınıflandırma algoritmalarının bir örnek üzerinde karşılaştırılması,” *inetd.org.tr*, Accessed: Nov. 09, 2021. [Online]. Available: https://www.inetd.org.tr/ab11/bildiri/67.pdf.

[3] B. ÖZLÜER BAŞER, M. YANGIN, and E. S. SARIDAŞ, “Makine Öğrenmesi Teknikleriyle Diyabet Hastalığının Sınıflandırılması,” *Süleyman Demirel Üniversitesi Fen Bilim. Enstitüsü Derg.*, Feb. 2021, doi: 10.19113/sdufenbed.842460.

[4] S. NAHZAT and M. YAĞANOĞLU, “Makine Öğrenimi Sınıflandırma Algoritmalarını Kullanarak Diyabet Tahmini,” *Eur. J. Sci. Technol.*, vol. 24, no. 24, pp. 53–59, Apr. 2021, doi: 10.31590/ejosat.899716.

[5] R. Birjais, A. Kumar Mourya, R. Chauhan, and H. Kaur, “Prediction and diagnosis of future diabetes risk: a machine learning approach,” 123AD, doi: 10.1007/s42452-019-1117-9.

[6] P. S. Kohli and S. Arora, “Application of machine learning in disease prediction,” *2018 4th Int. Conf. Comput. Commun. Autom. ICCCA 2018*, Dec. 2018, doi: 10.1109/CCAA.2018.8777449.

[7] S. K. Dey, A. Hossain, and M. M. Rahman, “Implementation of a Web Application to Predict Diabetes Disease: An Approach Using Machine Learning Algorithm,” *2018 21st Int. Conf. Comput. Inf. Technol. ICCIT 2018*, Jan. 2019, doi: 10.1109/ICCITECHN.2018.8631968.

[8] B. S. Murthy and J. Srilatha, “Comparative Analysis on Diabetes Dataset Using Machine Learning Algorithms,” *Proc. 6th Int. Conf. Commun. Electron. Syst. ICCES 2021*, pp. 1416–1422, Jul. 2021, doi: 10.1109/ICCES51350.2021.9488954.