NHANES Veri Seti ile Yaş ve Vücut Kitle İndeksi Tahmini

*Bilgisayar Mühendisliği, Fen Bilimleri Enstitüsü, Gazi Üniversitesi, Ankara, Türkiye* [*erbilbabacanoglu435@gmail.com.*](mailto:erbilbabacanoglu435@gmail.com.)

# ÖZET

Bu çalışmada, National Health and Nutrition Examination Survey (NHANES) veri seti kullanılarak yaş ve vücut kitle indeksi (BMI) tahmini gerçekleştirilmiştir. Farklı makine öğrenmesi algoritmalarının performanslarını karşılaştırmak amacıyla Naive Bayes, K-Nearest Neighbors (KNN) ve Karar Ağacı algoritmaları kullanılmıştır. Veriler üzerinde yapılan analizler sonucunda, KNN algoritması %95.26 doğruluk oranı ile en yüksek performansı göstermiştir. Naive Bayes algoritması %74.88 doğruluk oranı ile ortalama bir performans sergilemiştir. Karar Ağacı algoritması ise %94.04 doğruluk oranı ile yüksek bir performans göstermiştir.

Çalışmada, öncelikle veri seti temizlenmiş ve normalleştirilmiştir. Verilerin eksik ve hatalı olanları ayıklanmış ve ölçeklendirilmiştir. Ardından, veri seti eğitim ve test verilerine ayrılmış ve her bir algoritma üzerinde model eğitimi gerçekleştirilmiştir. Modellerin performansları doğruluk oranı, ortalama kare hatası ve öğrenme eğrisi analizleri ile değerlendirilmiştir. Naive Bayes, KNN ve Karar Ağacı algoritmaları ile yapılan tahminlerin sonuçları karşılaştırılmıştır.

Elde edilen sonuçlar, yaş ve BMI tahmininde KNN algoritmasının diğer algoritmalara göre daha yüksek doğruluk oranına sahip olduğunu ortaya koymuştur. KNN algoritmasının öğrenme eğrisi, modelin eğitim ve çapraz doğrulama skorlarının birbirine yakın olduğunu göstermiş ve bu da modelin genelleme yeteneğinin yüksek olduğunu ortaya koymuştur. Naive Bayes algoritması ise, özellikle düşük doğruluk oranı ile beklenen performansı gösterememiştir. Karar Ağacı algoritması, KNN algoritmasına yakın bir doğruluk oranı sunmuş, ancak aşırı uyum belirtileri göstermiştir.

**Anahtar Kelimeler:** Makine Öğrenmesi, Yaş Tahmini, Vücut Kitle İndeksi, Naive Bayes, K-Nearest Neighbors, Karar Ağacı.

# Giriş

Bu çalışmada, National Health and Nutrition Examination Survey (NHANES) veri seti kullanılarak yaş ve bmi tahmini gerçekleştirilmiştir. NHANES, Amerika Birleşik Devletleri'ndeki yetişkinler ve çocukların sağlık ve beslenme durumlarını değerlendirmek amacıyla oluşturulmuş kapsamlı bir veri setidir. Bu veri seti, bireylerin çeşitli sağlık göstergelerini ve demografik bilgilerini içermektedir. Yaş tahmini, özellikle yaşlı popülasyonlarda diyabet gibi kronik hastalıkların erken teşhisi ve tedavisi için büyük önem taşımaktadır. Yaş, sağlık risklerinin belirlenmesi ve uygun sağlık hizmetlerinin sağlanması açısından kritik bir değişkendir.

NHANES veri setindeki en önemli problemlerden biri, yaşın doğru tahmin edilmesidir. Doğru yaş tahmini, sağlık hizmetlerinin iyileştirilmesi ve hastalıkların önlenmesi açısından kritik öneme sahiptir. Bu çalışmada, NHANES veri seti kullanılarak yaş ve bmi tahmini için farklı makine öğrenmesi algoritmalarının performansları karşılaştırılmıştır. Çalışmada kullanılan yöntemler arasında K-Nearest Neighbors, Naive Bayes ve Karar Ağaçları yer almaktadır. Bu yöntemlerin her biri, farklı sonuçlara sahiptir ve sağlık verilerinin analizinde farklı sonuçlar verebilmektedir.

Bu çalışmada, yaş tahmini yapılırken LBXGLT (Glukoz Tolerans Testi), DIQ010 (Diyabet Teşhis Durumu) ve BMXBMI (Vücut Kitle İndeksi) gibi sağlık göstergeleri kullanılmıştır. Çalışmanın amacı, bu göstergeler kullanılarak yaş tahmininde en yüksek doğruluk oranına sahip makine öğrenmesi algoritmasını belirlemektir. Elde edilen sonuçlar, yaş tahmininde KNN algoritmasının diğer algoritmalara göre daha yüksek doğruluk oranına sahip olduğunu ortaya koymuştur. Naive Bayes ve Karar Ağaçları algoritma sonuçları ise KNN algoritmasına göre, düşük doğruluk oranlarla sonuçlanmıştır.

# Materyal ve Yöntem

* 1. **Veri Seti**

Bu çalışmada kullanılan veri seti, National Health and Nutrition Examination Survey (NHANES) veri setidir. NHANES, Amerika Birleşik Devletleri'ndeki yetişkinler ve çocukların sağlık ve beslenme durumlarını değerlendirmek amacıyla oluşturulmuş geniş kapsamlı bir veri setidir. Veri seti, Hastalık Kontrol ve Önleme Merkezleri aracılığıyla finanse edilmektedir. Veri setinin içeriği aşağıdaki gibidir.

* **RIDAGEYR (Yaş):** Katılımcıların yaşını ifade eder.
* **RIDRETH1 (Etnik Köken):** Katılımcıların etnik kökenini ifade eder.
* **RIAGENDR (Cinsiyet):** Katılımcıların cinsiyetini belirtir.
* **PAQ605 (Fiziksel Aktivite):** Katılımcıların fiziksel aktivite düzeyini belirtir.
* **BMXBMI (Vücut Kitle İndeksi):** Katılımcıların ağırlık ve boy ölçümlerinden hesaplanan bir indekstir.
* **LBXGLU (Glukoz):** Katılımcıların kan şekeri düzeylerini ölçer.
* **DIQ010 (Diyabet Durumu):** Katılımcıların diyabet durumunu gösterir.
* **LBXGLT (Glukoz Tolerans Testi):** Bireylerin kan şekeri düzeylerini ölçer.
* **LBXIN (İnsülin):** Katılımcıların insülin düzeylerini ölçer.

**Diyabetin tanımı ve önemi:**

Diyabet, vücudun insülin üretme veya kullanma yeteneğini etkileyen kronik bir hastalıktır. İnsülin, kan şekeri seviyelerini düzenlemeye yardımcı olan bir hormondur. Diyabetin üç ana tipi vardır: Tip 1, Tip 2 ve Gestasyonel Diyabet. Tip 2 diyabet, en yaygın diyabet türüdür ve genellikle yetişkinlerde görülür. Diyabet, kontrol edilmediğinde kalp hastalığı, böbrek hastalığı, sinir hasarı ve diğer ciddi sağlık sorunlarına yol açabilir. Bu nedenle, diyabetin erken teşhisi ve yönetimi, hastalığın komplikasyonlarını önlemek için kritik öneme sahiptir.

**Veri Setindeki Değerlerin Diyabeti Nasıl Etkilediği:**

Yaş, glukoz tolerans testi, vücut kitle indeksi ve diyabet durumu, diyabetin tanı ve yönetiminde önemli rol oynayan faktörlerdir. Bu değişkenler, diyabet riski ve yönetiminde kritik bilgiler sunar.

* **Yaş (RIDAGEYR):** Yaşlandıkça, vücutta insülin direnci artabilir ve diyabet gelişme riski yükselir.
* **Glukoz Tolerans Testi (LBXGLT):** Glukoz tolerans testi sonuçları, bireyin diyabet geliştirme riskini belirlemede kullanılır.
* **Vücut Kitle İndeksi (BMXBMI):** Yüksek BMI değerleri, obezite ile ilişkilidir ve obezite, diyabet gelişiminde önemli bir risk faktörüdür.
* **Diyabet Durumu (DIQ010):** Diyabet durumu, mevcut sağlık durumlarının değerlendirilmesinde kullanılır.

# Kullanılan Makine Öğrenmesi Algoritmaları

**Veri ön işleme**

NHANES veri seti, ham haliyle eksik ve hatalı veriler içerebilir. Bu nedenle, veri analizi ve model eğitimi öncesinde veri seti üzerinde çeşitli ön işleme adımları uygulanmıştır. İlk olarak, eksik veriler tespit edilmiştir. Özellikle, DIQ010 (Diyabet Durumu) değişkeni için "3" olarak işaretlenen eksik değerler, K-Nearest Neighbors (KNN) algoritması kullanılarak "1" (Diyabet Evet) veya "2" (Diyabet Hayır) olarak sınıflandırılmıştır. Bu işlem, eksik değerlerin giderilmesi ve veri setinin tamamlanması için önemlidir.

Ayrıca, veri setindeki bazı değişkenlerde ayrık değerler (outliers) tespit edilmiştir. Bu ayrık değerler, veri setindeki diğer değerlere kıyasla çok uçta yer almıştır. Ayrık değerlerin tespiti ve temizlenmesi, veri setinin güvenilirliğini artırmak ve model performansını iyileştirmek amacıyla gerçekleştirilmiştir. Sürekli değişkenler (örneğin, LBXGLT ve BMXBMI) normalleştirilmiştir. Normalleştirme, her bir özelliğin ortalama 0 ve standart sapma 1 olacak şekilde yeniden ölçeklendirilmesi işlemini içerir. Bu işlem, modelin daha hızlı ve etkili öğrenmesini sağlar.

**Veri Eğitimi**

Veri seti, eğitim ve test verileri olarak ikiye ayrılmıştır. Eğitim verisi, modelin öğrenmesi için kullanılan verileri içerirken, test verisi modelin performansını değerlendirmek için kullanılır. Bu çalışmada, veri seti %80 eğitim ve %20 test verisi olarak bölünmüştür. Ayrıca, modelin genelleme yeteneğini değerlendirmek için 5 katlı çapraz doğrulama (cross-validation) yöntemi kullanılmıştır. Bu yöntemde, veri seti beş eşit parçaya bölünmüş ve her bir parça sırasıyla test verisi olarak kullanılırken geri kalan parçalar eğitim verisi olarak kullanılmıştır. Bu süreç, modelin farklı veri alt kümeleri üzerinde nasıl performans gösterdiğini değerlendirir.

**Kullanılan Algoritmalar**

Bu çalışmada, yaş tahmini yapmak amacıyla üç farklı makine öğrenmesi algoritması kullanılmıştır: K-Nearest Neighbors (KNN), Karar Ağaçları ve Naive Bayes. Her bir algoritmanın temel prensipleri ve çalışma şekli aşağıda detaylı olarak açıklanmaktadır.

* K-Nearest Neighbors (KNN):

KNN algoritması, en yakın komşulara dayalı olarak sınıflandırma yapan basit ve etkili bir algoritmadır Bu çalışmada, k=5 olarak belirlenmiş ve model, glukoz tolerans testi vücut kitle endeksi özelliklerini kullanarak yaş tahmini yapmıştır.

* Karar Ağaçları:

Karar ağaçları, veri setini özelliklerine göre dallara ayırarak sınıflandırma yapan bir algoritmadır. Bu çalışmada, karar ağacı algoritması kullanılarak vücut kitle endeksi tahmini yapılmıştır

* Naive Bayes:

Naive Bayes algoritması, Bayes teoremini temel alan ve özellikle sınıflandırma problemlerinde kullanılan bir algoritmadır. Bu çalışmada, Naive Bayes algoritması kullanılarak yaş tahmini yapılmıştır

# Literatür Taraması

Diyabetin erken teşhisi ve yönetimi için makine öğrenimi tekniklerinin kullanımı son yıllarda büyük ilerlemeler kaydetmiştir. Birçok çalışma, National Health and Nutrition Examination Survey (NHANES) 2013-2014 veri setini kullanarak çeşitli makine öğrenimi algoritmalarının diyabet tahmininde yüksek doğruluk oranlarına ulaştığını göstermektedir.

Ko ve ekibi, XGBoost ve XGBoost Survival Embedding (XGBSE) modellerinin diyabet tahmininde yüksek performans sergilediğini ve uzun vadeli diyabet tahminlerinde yüksek doğruluk değerlerine ulaştığını göstermiştir. Başka bir araştırmada, GBM ve Logistic Regression teknikleri kullanılarak yapılan analizlerde, GBM modelinin %99.76 doğruluk oranı ile yüksek performans sergilediği gözlemlenmiştir. Yu ve arkadaşları, CATBoost, XGBoost, Random Forest, Logistic Regression ve SVM gibi beş farklı makine öğrenimi modelini karşılaştırmış ve CATBoost'un en yüksek doğruluğa sahip olduğunu bulmuşlardır. Aynı veri seti kullanılarak yapılan bir başka çalışmada ise, k-NN, SVM, Karar Ağaçları, Random Forest, AdaBoost ve Gaussian Naive Bayes gibi algoritmalar kullanılmış ve ensembıl modellerin, diyabet tahmininde tekli modellere göre daha iyi performans gösterdiği sonucuna varılmıştır. Ayrıca, çeşitli makine öğrenimi modellerinin (örneğin, Lojistik Regresyon ve SVM) diyabet tahminindeki performanslarını karşılaştıran diğer çalışmalarda da bu modellerin hastalık riski taşıyan bireylerin belirlenmesinde etkili olduğu bulunmuştur.

Bu çalışmalar, makine öğrenimi algoritmalarının diyabet tahmininde etkin bir şekilde kullanılabileceğini ve doğru model ve veri işleme teknikleri ile bu modellerin doğruluk oranlarının artırılabileceğini göstermektedir. Bu bağlamda, diyabet tahmini ve yönetimi için makine öğrenimi algoritmalarının potansiyeli büyüktür ve gelecekteki çalışmalar bu alandaki uygulamaların daha da geliştirilmesine katkı sağlayacaktır.

# Bulgular

Bu çalışmada elde edilen bulgular, yaş tahmini için KNN algoritmasının diğer algoritmalara göre daha yüksek doğruluk oranına sahip olduğunu göstermektedir. ncak, Naive Bayes algoritması, özellikle bu veri seti üzerinde düşük performans göstermiştir. Elde edilen sonuçlar, KNN algoritmasının yaş tahmini için daha uygun olduğunu ve Karar Ağaçları algoritmasının da BMI tahmininde güçlü bir alternatif olduğu gözlemlenmiştir. Aşağıda her modelin bulgusu açıklamalarından sonra şekliyle verilmiştir.

* K-Nearest Neighbors (KNN) ile yaş tahmini

KNN algoritması, glukoz tolerans testi ve vücut kitle endeksi özelliklerini kullanarak yaş tahmini yapmak için kullanılmıştır. KNN algoritmasının doğruluk oranı %95.26 olarak hesaplanmıştır. Modelin öğrenme eğrisi, eğitim ve çapraz doğrulama skorlarının birbirine yakın olduğunu ve modelin genelleme yeteneğinin yüksek olduğunu göstermektedir. Aşağıda bulguya dair şekil

metin, ekran görüntüsü, renklilik içeren bir resim

Açıklama otomatik olarak oluşturuldu

*Şekil 1. KNN ile yaş tahmini*

* Karar ağacı ile vücut kitle endeksi tahmini

Karar Ağaçları algoritması, yaş ve glukoz tolerans testi özelliklerini kullanarak vücut kitle indeksi tahmini yapmak için kullanılmıştır. Karar Ağacı algoritmasının doğruluk oranı %94.04 olarak hesaplanmıştır. Modelin öğrenme eğrisi, eğitim ve test skorları arasında belirgin bir fark olmadığını ve modelin aşırı uyum yapmadığını göstermektedir.

metin, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

*Şekil 2. Karar Ağacı ile vücut kitle endeksi tahmini*

* Naive Bayes ile Yaş Tahmini

Naive Bayes algoritması, glukoz tolerans testi ve glukoz özelliklerini kullanarak yaş tahmini yapmak için kullanılmıştır. Naive Bayes algoritmasının doğruluk oranı %74.88 olarak hesaplanmıştır. Modelin öğrenme eğrisi, eğitim ve çapraz doğrulama skorları arasında büyük bir fark olduğunu ve modelin diğer kullanılan modellere göre düşük performans gösterdiğini ortaya koymaktadır.

metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Açıklama otomatik olarak oluşturuldu

*Şekil 3. Naive Bayes ile yaş tahmini*

# Sonuç ve Öneriler

Bu çalışmada, National Health and Nutrition Examination Survey (NHANES) veri seti kullanılarak yaş ve vücut kitle indeksi tahmini için üç farklı makine öğrenmesi algoritması uygulanmıştır.

Elde edilen bulgular üzerine, KNN algoritması, %95.26 doğruluk oranı ile yaş tahmininde en başarılı algoritma olarak belirlenmiştir. Bu sonuç, KNN'nin glukoz tolerans testi ve vücut kitle indeksi gibi sağlık göstergeleri kullanılarak yaş tahmini yapmada etkili bir yöntem olduğunu göstermektedir. Karar Ağaçları algoritması, %94.04 doğruluk oranı ile BMI tahmininde yüksek bir performans sergilemiş ve sağlık verilerinin görselleştirilmesi ve anlaşılmasında önemli bir araç olmuştur. Naive Bayes algoritması ise %74.88 doğruluk oranı ile diğer modellerin performanslarının altında bir performans göstermiştir.

**Öneriler:**

* **Genişletilmiş Veri Setleri:** Gelecekteki çalışmalarda, daha geniş veri setleri kullanılarak modellerin performansları değerlendirilebilir. Daha fazla veri, modellerin genelleme yeteneğini artırabilir ve daha güvenilir sonuçlar elde edilmesine olanak tanır.
* **Farklı Makine Öğrenmesi Algoritmaları:** Diğer makine öğrenmesi algoritmalarının kullanılması ve karşılaştırılması, daha iyi performans sonuçlarına yol açabilir.
* **Çoklu Özellik Kullanımı:** Sadece yaş ve BMI tahmini yerine, diğer sağlık göstergelerinin de dahil edilmesiyle daha kapsamlı analizler yapılabilir. Bu, modellerin performansını ve uygulanabilirliğini artırabilir.

Tablo 1. Elde edilen sonuçlar

|  |  |
| --- | --- |
| Değerlendirme Kriterleri | NHANES Veri Seti Doğruluk Tahmin Oranı |
| K-En Yakın Komşu | **0,9526** |
| Karar Ağacı | 0,9404 |
| Basit Bayes | 0,7488 |

# KAYNAKÇA

* <https://www.mdpi.com/1660-4601/19/19/12378>
* <https://bmcendocrdisord.biomedcentral.com/articles/10.1186/s12902-019-0436-6>
* <https://www.researchsquare.com/article/rs-3764619/v1>
* <https://www.sciencedirect.com/science/article/abs/pii/S0168822719309404>
* <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-023-05465-z>
* <https://www.datasciencecentral.com/>
* <https://chatgpt.com/>
* <https://github.com/>
* <https://www.dataquest.io/>