Penguen Türlerinin Morfolojik Özelliklerine Dayalı Kümeleme ve Cinsiyet Tahmini: Veri Odaklı Yaklaşımlar ve Makine Öğrenimi Teknikleri

Emine AKDENİZ-202013171052, İrem Yağmur BARDAK-202013171074

Özet

Bu çalışmada, penguenlerin morfolojik özelliklerine dayanarak farklı türlerin gruplandırılması ve analiz edilmesi amaçlanmıştır. Veriler Kaggle platformundan alınan "Clustering Penguins Species" adlı veri setinden elde edilmiştir. Çalışma, penguenlerin culmen uzunluğu ve derinliği, yüzgeç uzunluğu, vücut ağırlığı gibi özelliklerini kullanarak türlerin ekolojik adaptasyonlarını ve beslenme stratejilerini anlamaya odaklanmıştır.Veri seti temizlenmiş ve standartlaştırılmıştır. Temel Bileşen Analizi (PCA) ile veri setinin boyutu azaltılarak %90 varyansı açıklayacak bileşenler belirlenmiştir. K-means ve hiyerarşik kümeleme algoritmaları kullanılarak penguen türleri gruplandırılmış ve optimal küme sayısı Dirsek Yöntemi ile belirlenmiştir.LVQ (Learning Vector Quantization) modeli ve ANFIS (Adaptive Neuro-Fuzzy Inference System) modeli kullanılarak penguen cinsiyeti tahmin edilmiştir. Bu modellerin performansı çeşitli metriklerle değerlendirilmiş ve sonuçlar görselleştirilmiştir. %90 doğruluk ile LVQ'nün bu veri seti için daha uygun olduğu ortaya konmuştur. Elde edilen bulgular, penguen türlerinin ekolojik nişlerini ve evrimsel ilişkilerini anlamak için önemli bilgiler sunmakta ve penguen popülasyonlarını koruma stratejileri geliştirmede rehberlik etmektedir. Bu çalışma, veri odaklı yaklaşımlar ve makine öğrenimi tekniklerinin biyolojik araştırmalarda ve koruma çalışmalarında nasıl kullanılabileceğine dair değerli bir örnek teşkil etmektedir.

Abstract

This study aims to group and analyze different species of penguins based on their morphological characteristics. The data was obtained from the dataset named "Clustering Penguins Species" from the Kaggle platform. The study focused on understanding the ecological adaptations and feeding strategies of the species using characteristics such as culmen length and depth, fin length, body weight, etc. The dataset was cleaned and standardized. Principal Component Analysis (PCA) was used to reduce the size of the dataset and identify components that explain 90% of the variance. Penguin species were grouped using K-means and hierarchical clustering algorithms and the optimal number of clusters was determined by the Elbow Method.Penguin sex was predicted using the LVQ (Learning Vector Quantization) model and the ANFIS (Adaptive Neuro-Fuzzy Inference System) model. The performance of these models was evaluated with various metrics and the results were visualized. It was revealed that LVQ was more suitable for this data set with 90% accuracy. The findings provide important insights for understanding the ecological niches and evolutionary relationships of penguin species and provide guidance for developing strategies to conserve penguin populations. This study provides a valuable example of how data-driven approaches and machine learning techniques can be used in biological research and conservation efforts.

1.Giriş

Günümüzde, doğa ve ekosistemler üzerindeki etkilerimizi anlamak ve korumak için veri odaklı yaklaşımlar giderek daha fazla önem kazanıyor. Penguenler, kutup bölgelerinde ve güney yarımküredeki soğuk denizlerde yaşayan uçamayan deniz kuşlarıdır. Morfolojik özellikleri, bu kuşların yaşadığı ortama olan adaptasyonlarını ve farklı türler arasındaki çeşitliliği yansıtır [1].

Culmen uzunluğu ve derinliği, penguen türlerinin beslenme alışkanlıklarını ve avlanma stratejilerini anlamak için önemlidir. Örneğin, daha uzun ve derin bir culmene sahip olan penguenler genellikle derin sularda avlanırken, daha kısa ve sivri culmene sahip olanlar genellikle yüzeyde avlanır.

Penguenlerin yüzgeçleri, su altında hareket etmelerini sağlayan önemli bir uzuvlarıdır. Yüzgeç uzunluğu, penguenlerin su altında manevra kabiliyetlerini ve yüzme yeteneklerini etkiler. Bazı türlerde yüzgeçler daha uzun ve daha genişken, diğerlerinde daha kısa ve dar olabilir. Bu özellik, penguen türlerinin avlanma stratejileri ve su altında hareket kabiliyetlerini belirler.

Vücut ağırlığı, penguenlerin adaptasyonları ve yaşadıkları ortama uyum sağlama yetenekleri ile ilişkilidir. Daha büyük vücut ağırlığına sahip penguenler genellikle daha derin sularda ve daha uzun süre dalmayı tercih ederken, daha hafif olanlar genellikle yüzeyde daha fazla zaman geçirir.

Cinsiyet, penguen popülasyonlarının demografik yapısını ve üreme davranışlarını anlamak için kritik bir faktördür. Bazı penguen türlerinde, erkek ve dişi bireyler arasında görsel veya fiziksel farklılıklar bulunabilir. Örneğin, bazı türlerde erkeklerin daha büyük boyutlarda ve daha canlı renklere sahip olduğu görülür [2,3].

Penguen türleri arasındaki bu morfolojik farklılıklar, her bir türün yaşadığı çevreye özgü adaptasyonlarını yansıtır. Örneğin, Antarktika'da yaşayan İmparator Penguenler, büyük vücutları ve uzun yüzgeçleriyle derin sularda uzun süre dalmaya uygunken, Galapagos Adaları'nda yaşayan Galapagos Penguenleri, daha küçük boyutları ve kısa yüzgeçleriyle daha sığ sularda avlanmaya uyum sağlamıştır. Bu adaptasyonlar, penguen türlerinin hayatta kalma ve üreme başarısını etkiler.

Denetimsiz kümeleme algoritmaları aracılığıyla, penguen gruplarını fiziksel özelliklerine göre tanımlayabiliriz; bu da onların evrimsel ayrışmasına ve ekolojik nişlerine dair anlayışlar sağlar ve dünya çapındaki benzersiz penguen türlerini korumaya yönelik hedeflenmiş koruma çabalarına yol gösterir. Kümeleme algoritmalarının gücünden faydalanarak, penguen popülasyonu içindeki gizli desenleri ve ilişkileri ortaya çıkarabiliriz, sonuç olarak her penguen türünün benzersiz özelliklerini ve adaptasyonlarını hesaba katan daha etkili koruma stratejileri oluşturabiliriz. Penguen türlerini ölçülen özelliklerine dayanarak gruplandırmak için k-ortalama ve hiyerarşik kümeleme tekniklerini kullanacağız ve penguen popülasyonunun temel yapısını ortaya çıkaracağız. Penguen türlerini morfolojik özelliklerine göre kümeleyerek, evrimsel ilişkilerini ve ekolojik özelleşmelerini belirleyebiliriz, benzersiz alt grupları tanımlayabiliriz ve her penguen türü için özelleştirilmiş koruma planları geliştirebiliriz [4].

2. Literatür Çalışması

Zhang ve diğ.(2015) Çalışmaların da hayvan davranışlarını coğrafi yaşam çizgileri içinde sınıflandırmışlardır. Davranışsal Değişim Noktası Analizi, Hiyerarşik Çok Değişkenli Küme Analizi ve K-Ortalama Kümeleme yöntemleri kullanılmıştır. Sonuçlar, modelleme prosedürünün, sentetik yörüngelerdeki tüm bireysel konum gözlemlerinin %92,5'ini

doğru bir şekilde sınıflandırdığını ve davranış modlarını başarılı bir şekilde ayırt etme konusunda makul bir yetenek gösterdiğini göstermektedir [5].

Gemma ve diğ. (2018), beş penguen türünün üreme kolonileri arasındaki bağlantı desenlerini karşılaştırmak için soydaş, örtüşen dağılımlara sahip ve farklı ekolojik nişlere sahip olan bir karşılaştırmalı çerçeve ve RAD-Seq aracılığıyla elde edilen genom geniş veriler kullanmıştır. Bulgular, pelajik ve kıyısal nişlere sahip penguen türlerinin genetik farklılaşma desenlerini belirleyen önemli faktörleri ortaya koymaktadır. Pelajik türler genellikle genetik olarak benzerken, kıyısal nişe sahip gentoo penguenleri yüksek düzeyde genetik farklılaşma göstermektedir. Denizdeki yaşam alanının dispersal desenlerini belirlemede en önemli faktör olduğu bulgusu, pelajik türlerin gen akışını kolaylaştırdığını göstermektedir. Buna karşılık, kıyısal nişe sahip penguenlerin dağılımını sınırlayan faktörler bulunmaktadır. Bu bulgular, nesli tükenme riski tahminlerinde ve koruma stratejilerinde dispersalın önemini vurgular[6].

Jane ve diğ.(2017),imparator penguenlerinin popülasyon yapısını belirlemek için geniş kapsamlı bir analiz yapmaktadır. Önceki çalışmaların aksine, Antarktika çevresindeki sekiz farklı koloniden elde edilen veriler kullanılarak yapılan analizler, imparator penguenlerinin en az dört farklı metapopülasyondan oluştuğunu ve Ross Denizi'nin net bir şekilde ayrı bir metapopülasyon olduğunu göstermektedir. Bu bulgular, türün etkili korunması için popülasyon yapısının anlaşılmasının önemini vurgulamaktadır. Geniş SNP veri setleri kullanarak belirgin olmayan popülasyon yapılarını tespit etme ve yorumlama konusundaki zorluklar da tartışılmaktadır. Sonuç olarak, tehdit altındaki türler için yönetim stratejilerinin belirlenmesinde bu tür yapıların dikkate alınması gerektiği vurgulanmaktadır [7].

Piotr ve diğ.(2013), Palaeeudyptes penguenlerinin Eosen Antarktika temsilcileri içindeki morfometrik desenleri ortaya koyan, geliştirilmiş kümeleme teknikleri, temel koordinatlar yöntemi, çekirdek yoğunluk tahminleri ve sonlu karışım modeli analizleri gibi çoklu ve tekli veri analizlerinin sonuçlarını rapor ediyoruz. Bu büyük boyutlu kuşlar, çoğunlukla birçok izole kemikten bilinen iki tür olan P. gunnari ve P. klekowskii tarafından temsil edilmiştir. Araştırmalar, erken penguenlerin paleontolojisinde önemli kemikler olan tarsometatarsi'ye odaklanmış olup, incelenen örneklerin kısmen örtüşen boyut dağılımlarına sahip iki taksona "bulanık" bölütleme imkanı tanıyan olasılığa dayalı bir çerçeve sonuç vermiştir. Bu kadar çok sayıda tür, hem çoklu hem de tekli çalışmalardan elde edilen sonuçlarla desteklenmiştir. Bizce, Antarktika'daki Palaeeudyptes türleri arasındaki ayırımı yaparken formun nicel analizine daha fazla güvenilmesi gerekmektedir [8].

Jason ve diğ. (1998) Bu çalışmada, dalış verilerini sınıflandırmak için çeşitli algoritmaları incelemekte ve değerlendirmektedir. Bunlar, istatistik alanından k-means ve fuzzy c-means kümeleme teknikleri ile yapay sinir ağları alanından Kohonen öz-düzenleyen harita (SOM) ve fuzzy uyumlu rezonans teorisi (ART) tekniklerini içermektedir. Monte Carlo simulasyonu, bilinen çözümlere sahip yapay olarak üretilen veriler üzerinde farklı koşullar altında kümeleme performansını test etmek için gerçekleştirilmiştir. K-means, fuzzy c-means ve SOM, yapay olarak üretilen veriler üzerinde eşit derecede iyi performans gösterirken, fuzzy ART'nin hata oranları iki kat daha yüksekti [9].

Hart ve diğ (2010) Bu çalışmada, bu sorunu çözmek için gizli bir Markov modeli (HMM) adlı bir makine öğrenme tekniği kullanılmıştır. HMM, bir sistemin temel durumlarını gözlemlenebilir çıktılardan tanımlamayı amaçlar. Bu

çalışmada, Bird Adası'nda üreyen 103 makarna pengueninin dalış verileri kullanılarak HMM uygulanmıştır. Penguenlerin iki farklı davranış sınıfı belirlenmiştir: kısa-sığ ve uzun-derin dalışlar. Bu iki davranış sınıfı kullanılarak, bu durumlar arasındaki geçiş olasılıkları hesaplanmış ve bu geçiş olasılıklarındaki değişikliklerin neyi öngördüğü analiz edilmiştir. Araştırma sonuçları, üreme aşamasının, bireyin cinsiyetinin ve yılının, uzun-derin ve kısa-sığ sıralı dalışlar arasındaki geçiş olasılığını etkilediğini göstermiştir. Ayrıca, günlük bir döngü boyunca dört üreme aşaması arasındaki saatlik geçiş oranlarında da farklılıklar belirlenmiştir [10].

Chessa ve diğ. (2017) Bu çalışmada, penguen ivmeölçer verilerinden av yakalama olaylarını tanımlamak için Destek Vektör Makinesi (SVM) ve Giriş Gecikmeli Sinir Ağı (IDNN) modelleri arasında karşılaştırmalı bir analiz sunulmaktadır. Arkaya monteli ivmeölçerlerden alınan 3 boyutlu zaman serisi verileri önceden sınıflandırılmış bir veri seti kullanılarak incelenmiştir. Her iki model de penguenlerin davranışlarını belirli aralıklarla 'avı idare etme' veya 'yüzme' olarak sınıflandırmak için eğitilmiştir. IDNN modelinin, SVM ile aynı düzeyde sınıflandırma doğruluğuna, ancak daha az bellek gereksinimiyle ulaşıp ulaşamayacağı belirlenmeye çalışılmıştır. Deneysel sonuçlar, her iki modelin de yaklaşık olarak eşdeğer bir doğruluk elde ettiğini göstermektedir. IDNN için 0,5 kB ve SVM için 0,7 Mb bellek talebiyle öne çıkan verileri %85 kullanıyor. Ham ivmeölçer verileri, modellerin genelleştirilebilirliğini biraz daha düşük bir doğrulukla yaklaşık %80'e çıkarmamıza olanak tanıyor. Bu, IDNN modelinin ivme ölçerin kendisine yerleştirilebileceğini ve ham zaman serisi verilerinin alınması ve kaybıyla ilgili sorunları azaltabileceğini gösterir [11].

Le ve diğ (2019) Bu çalışmada, yüksek çözünürlüklü uydu görüntülerinde Adélie penguen kolonilerinin anlamsal segmentasyonunu gerçekleştirmek için derin öğrenme tabanlı bir model sunulmaktadır. Segmentasyon modellerinin eğitimi için Antarktika'daki 193 Adélie penguen kolonisinden 2044 jeoreferanslı kırpılmış görüntüler içeren benzersiz bir veri kümesi olan Penguin Colony Dataset kullanılmıştır. Piksel düzeyinde etiket maskelerinin yetersizliği nedeniyle, zayıf denetimli bir çerçeve önerilmiş ve zayıf etiketlerden etkin bir şekilde bir segmentasyon modeli öğrenilmiştir. Segmentasyon ağı, zayıf etiketli verilerden öğrenmek için ortalama aktivasyona dayalı özel bir kayıp fonksiyonu kullanılarak eğitilmiştir. Deneyler, zayıf etiketli eğitim örneklerinin eklenmesinin segmentasyon performansını önemli ölçüde artırdığını göstermiştir; Penguin Colony Dataset üzerinde Ortalama Birleşim-üzeri Birlik'in ortalama yüzdesini %42.3'ten %60.0'a çıkarmıştır [12].

3. Materyal ve Metod

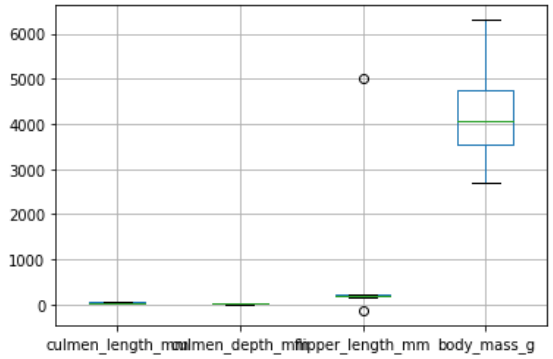
3.1 Veri Seti

Veri seti, Kaggle platformundan alınmıştır ve "Clustering Penguins Species" adlı veri setidir. 345 örneğe sahip bir veri seti bulunmaktadır. Bu veri seti 5 sütuna sahiptir ve her bir örnek tek bir etikete sahiptir. Veri setindeki eksik veriler temizlenmiş ve olası anlamsız veya hatalı değerler filtrelenmiştir. Daha sonra, denetimsiz öğrenme algoritmalarında kullanılmak üzere cinsiyet bilgisini içermeyen bir veri çerçevesi oluşturulmuştur. Son olarak, veriler standart hale dönüştürülmüş, bu da analiz ve modelleme süreçlerinde tutarlılık sağlamaktadır [13].

culmen_length_mm	Külmen Uzunluğu (mm)
culmen_depth_mm	Külmen Derinliği (mm)
flipper_length_mm	Yüzgeç Uzunluğu (mm)

body_mass_g	Vücut Ağırlığı (g)
sex	Cinsiyet

Tablo 1 V*eri Setinin Özellikleri*



Tablo 2 *Veri Setinin Dağılım Kutu Grafiği*

3.2 Sınıflandırma Yöntemleri

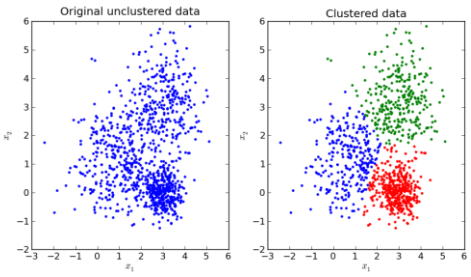
3.2.1 K-Ortalama (K-Means)

K-means, veri analitiğinde sıkça kullanılan bir kümeleme algoritmasıdır. Bu algoritma, veri noktalarını belirli bir sayıda kümeye böler ve her veri noktasını en yakın kümeye atar. Bu temel prensip, veri noktalarını kümeler arasında homojenlik ve kümelerin merkezlerine olan uzaklığın minimize edilmesi üzerine kuruludur.

Algoritma, belirli adımları takip eder. İlk olarak, kullanıcı tarafından belirlenen kümelerin sayısı (k) belirlenir ve her bir kümeye rastgele merkezler atanır. Daha sonra, her veri noktası, bu kümeler arasında mesafesine göre en yakın kümeye atanır. Ardından, her veri noktası atandığı kümeye göre bir aritmetik ortalama hesaplanarak, kümelerin yeni merkezleri belirlenir.

Bu işlem, kümelerin merkezlerinin değişmeyene kadar veya belirli bir iterasyon sayısına ulaşılana kadar tekrarlanır. Sonuç olarak, algoritma kümelerin merkezlerinin artık değişmediği veya belirli bir konverjans kriterine ulaşıldığı noktada durur.

K-means, kümeleme problemlerinde yaygın olarak kullanılmaktadır çünkü basit ve etkilidir. Ancak, başlangıç merkezlerinin rastgele seçilmesi, algoritmanın başarımını etkileyebilir ve farklı başlangıç noktaları farklı sonuçlara yol açabilir. Bu nedenle, genellikle algoritmanın sonuçlarına güvenmek için birden fazla kez çalıştırılır ve en iyi sonuçlar seçilir. Bu özellikleriyle, K-means, veri analitiği alanında önemli bir araç olarak kabul edilir [14,15].



Şekil 1 *K-Means Kümeleme Algoritması*

3.2.2 Öğrenme Vektör Kuantizasyonu (Learning Vector Quantization-LVQ)

Öğrenen Vektör Kuantizasyonu (LVQ), desen tanıma ve makine öğrenimi alanlarında yaygın olarak kullanılan bir denetimli öğrenme algoritmasıdır. LVQ, sınıflandırma problemlerini çözmek için tasarlanmıştır ve veri kümesindeki örnekler arasındaki ilişkileri modellemek için etkili bir yaklaşım sunar.

Başlangıç ağırlıklarının belirlenmesiyle başlayan LVQ'da, her sınıfa bir veya birden fazla ağırlık vektörü atanır. Bu ağırlık vektörleri, farklı sınıflara ait örneklerin temsil edilmesi için kullanılır ve modelin başlangıç durumunu tanımlar.

Her bir giriş örneği, ağırlık vektörlerine olan benzerliğine göre sınıflandırılır. Mesafe ölçümü kullanılarak, giriş örneği ile ağırlık vektörleri arasındaki ilişki belirlenir ve en yakın ağırlık vektörüne ait sınıf atanır. Bu adım, mevcut ağırlık vektörlerinin kullanılmasıyla sınıflandırma işlemini gerçekleştirir.

Sınıflandırma sonucunda yanlış sınıflandırılan örnekler için ağırlık vektörleri güncellenir. Yanlış sınıflandırılan örneğin özellikleri ile en yakın ağırlık vektörü arasındaki ilişkiye bağlı olarak güncelleme yapılır. Bu sayede, model daha iyi sınıflandırma yapacak şekilde ayarlanır.

Sınıflandırma hataları düzeltilene kadar sınıflandırma ve ağırlık vektörlerinin güncellenmesi adımları tekrarlanır. Bu döngüsel süreç, algoritmanın örnekler arasındaki ilişkileri daha iyi öğrenmesini sağlar ve model her iterasyonda daha iyi hale gelir.

LVQ algoritması, belirli sınıflara ait örneklerin sınıflandırılması için oldukça etkilidir. Ancak, büyük veri setleri veya karmaşık yapılarla başa çıkmak için tek başına yeterli olmayabilir. Bu durumlarda, daha karmaşık algoritmalar veya özellik mühendisliği gibi tekniklerle birlikte kullanılabilir [16,17].

3.2.3 Uyarlanabilir Nöro-Bulanık Çıkarım Sistemi (Adaptive Neuro-Fuzzy Inference System-ANFIS)

ANFIS (Adaptive Neuro-Fuzzy Inference System), karmaşık sistemlerin modellemesi için etkili bir makine öğrenimi yöntemidir, çünkü yapay sinir ağları (ANN) ve bulanık mantık (FL) tekniklerini birleştirir. Bu hibrid sistem, karmaşık ilişkileri modelleme yeteneği sağlar. ANFIS, beş katmandan oluşur. İlk katman olan giriş katmanında, sisteme girdi sağlayan değişkenler bulunur ve her giriş değişkeni bir düğümle temsil edilir.

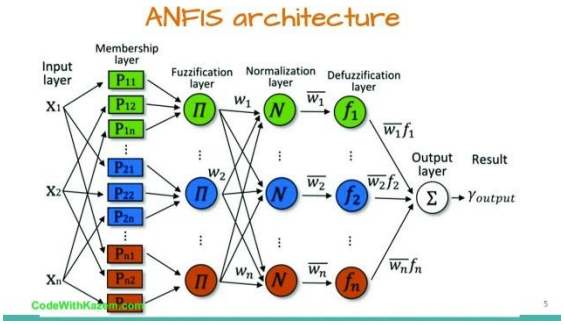
Bulanıklaştırma katmanı, giriş değişkenlerini bulanıklaştırır ve bu aşamada girdiler belirli bir bulanıklık terimine dönüştürülür.

Karmaşıklık katmanı, bulanıklaştırılmış girdileri birbirleriyle çarparak karmaşık terimler oluşturur ve bu adım, girdiler arasındaki ilişkileri daha da karmaşık hale getirir.

Birleştirme katmanı, karmaşık terimleri toplayarak birleştirir ve çıkış katmanına doğru iletilir. Bu katmanda, sistemdeki karmaşıklık azaltılır ve daha özgül çıktılar elde edilir.Son olarak, çıkış katmanında sistemden bir çıkış elde edilir ve burada bir çıkış değişkeni için bir çıkış değeri üretilir.

ANFIS'in eğitim süreci, belirli bir veri setine göre ayarlanır ve bu süreçte ağırlıklar ve parametreler optimize edilir. Veriler ANFIS'e sunulduğunda, giriş değişkenlerinin bulanıklaştırılması, karmaşık terimlerin hesaplanması ve çıkış değerlerinin belirlenmesi gibi adımlar izlenir.

ANFIS'in çalışma prensibi, bulanık mantığın esnekliğinden ve yapay sinir ağlarının öğrenme yeteneklerinden yararlanır. Bu sayede, özellikle belirsizlik içeren ve karmaşık sistemlerin modellemesi gereken durumlarda etkili bir yöntem olarak öne çıkar [18,19].



Şekil 2 *ANFIS Mimarisi*

3.2.4 Yapay Sinir Ağları (Artificial Neural Network-ANN)

Yapay sinir ağları (YSA'lar) veya simüle edilmiş sinir ağları (SNN'ler) olarak da adlandırılan sinir ağları, makine öğreniminin öne çıkan bir alt kümesidir ve derin öğrenme algoritmalarının temelini oluşturur. Bu ağlar, isimleri ve yapılarıyla insan beynindeki biyolojik nöronların etkileşimini taklit ederek tasarlanmıştır. Yapay sinir ağları, genellikle bir girdi

katmanı, bir veya daha fazla gizli katman ve bir çıktı katmanından oluşan düğüm katmanlarından meydana gelir. Her düğüm veya yapay nöron, diğer düğümlerle bağlantılıdır ve bu bağlantılar ağırlıklar ve eşik değerleri ile belirlenir. Bir düğüm, çıktısı belirlenen eşik değerinin üzerindeyse etkinleşir ve ağın bir sonraki katmanına veri aktarır; aksi takdirde, veri aktarımı gerçekleşmez. Sinir ağları, eğitim verilerine dayanarak zaman içinde doğruluklarını öğrenir ve geliştirir. Bu öğrenme süreci, bilgisayar bilimi ve yapay zeka alanında güçlü araçlar sunar. Özellikle eğitim sonrasında, yüksek hızda veri sınıflandırma ve kümeleme yetenekleri sağlar. Örneğin, konuşma tanıma veya görüntü tanıma gibi görevler, manuel tanımlamalara kıyasla çok daha hızlı bir şekilde gerçekleştirilebilir. Bu nedenle, sinir ağları, Google'ın arama algoritması gibi büyük ve karmaşık görevlerde başarıyla kullanılan güçlü araçlardan biridir [20,21].

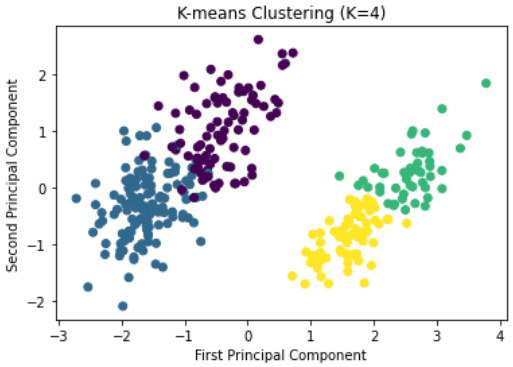
4. Bulgular ve Sonuç

K-Means kümeleme algorştmasında ilk olarak, veri seti bir CSV dosyasından yüklenmiş ve özelliklerinin dağılımını görselleştirmek için bir kutu grafiği oluşturulmuştur. Ardından, veri ön işleme adımında eksik veya geçersiz değerlere sahip satırlar atılmış ve "flipper_length_mm" sütunundaki aykırı değerlerin çıkarılması için bir aralık filtresi uygulanmıştır. Kategorik değişkenler tek-ısır kodlanarak sayısal hale getirilmiştir.

Daha sonra, veri StandartScaler kullanılarak sıfır ortalamaya ve birim varyansa sahip olacak şekilde standart hale getirilmiştir. Bu adımdan sonra, Temel Bileşen Analizi (PCA) uygulanarak veri setinin boyutu azaltılmış ve %90 varyansı açıklayacak bileşen sayısı belirlenmiştir.

Optimal küme sayısını belirlemek için azaltılmış veri kümesi üzerinde Dirsek Yöntemi kullanılmıştır. Daha sonra, belirlenen optimal küme sayısı ile K-means kümeleme gerçekleştirilmiştir. Elde edilen kümeler, ilk iki temel bileşen kullanılarak bir saçılma grafiğinde görselleştirilmiştir.

Son olarak, farklı başlangıç yöntemleri (k-means++, random ve PCA tabanlı) için K-means kümeleme yöntemleri, eğilim, homojenlik, tamamlama, V-ölçümü, uyarlanmış Rand endeksi, uyarlanmış karşılıklı bilgi ve siluet puanı gibi çeşitli ölçütlere göre karşılaştırılmıştır. Bu yöntemlerin performansı, çeşitli kümeleme ölçütleri kullanılarak değerlendirilmiştir.



Şekil 3 *K-Means Görselleştirme*

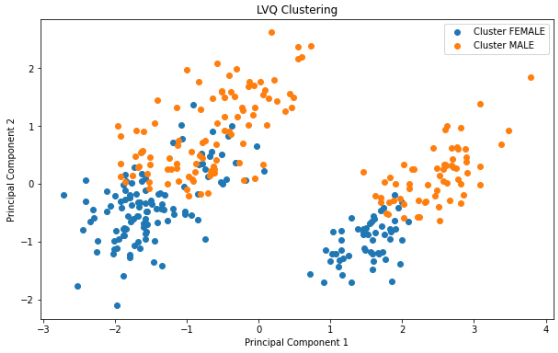
LVQ Modeli, belirli sayıda prototip vektörle temsil edilen bir model oluşturur. Prototipler, veri noktaları arasındaki uzaklığı minimize etmek için eğitilir. Model, veri noktalarını en yakın prototip vektöre atayarak küme tahminlerini yapar. Bu tahminler gerçek etiketlerle karşılaştırılarak doğruluk değeri hesaplanır.

PCA kullanılarak veri boyutu azaltılır. Veri, temel bileşenlere dönüştürülerek daha düşük boyutlu bir uzayda temsil edilir. Bu, veri setinin karmaşıklığını azaltarak işleme süresini iyileştirir ve model performansını artırır.

Elde edilen kümeleme sonuçları, iki ana bileşeni gösteren bir grafik üzerinde görselleştirilir. Her bir küme farklı renklerle gösterilir. Bu, veri setindeki yapıları ve benzerlikleri daha anlaşılır bir şekilde görselleştirmeye yardımcı olur. Görselleştirmeler, veri analizinde önemli

bir araçtır ve karmaşık yapıları anlamak için kullanılır.

Bu çalışma, LVQ algoritması kullanarak veri setinin derinlemesine analizini gerçekleştirmiştir. Elde edilen sonuçlar, veri setindeki yapıları ve ilişkileri daha iyi anlamamıza yardımcı olur. Bu bilgi, karar verme süreçlerinde ve stratejik planlamada değerli bir kaynak olabilir. Gelecekteki çalışmalar, farklı veri setleri üzerinde LVQ'nun performansını ve uygulanabilirliğini daha da inceleyebilir.



Şekil 4 *LVQ Cinsiyet Dağılımı*

ANFIS modeli, Adaptive Neuro-Fuzzy Inference System'ın kısaltmasıdır ve bulanık mantık temelli bir yapay zeka modelidir. Model oluşturulurken, giriş değişkenleri (culmen_length_mm, culmen_depth_mm, flipper_length_mm, body_mass_g) ve çıkış değişkeni (sex) tanımlanır.

Üyelik fonksiyonları, her bir giriş değişkeni için tanımlanır. Bu fonksiyonlar, her değişkenin belirli bir duruma veya kategoriye ait olma olasılığını ifade eder.

Sonrasında, kurallar belirlenir. Örneğin, "kısa culmen_length veya sığ culmen_depth veya kısa flipper_length veya hafif body_mass ise cinsiyet dişi olabilir" gibi kurallar modelin çıkarım sürecini yönlendirir

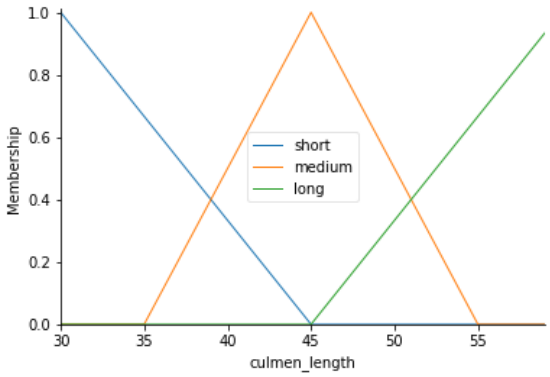
Bulanık mantık sistemi oluşturulur ve bir kontrol sistemi simülasyonu başlatılır. Bu simülasyon, belirlenen kurallar ve üyelik fonksiyonları doğrultusunda verilerin çıkarımını gerçekleştirir.

Model eğitimi aşamasında, her bir penguen örneği için girişler belirlenir, model tarafından çıktı tahmin edilir ve gerçek cinsiyetle karşılaştırılır. Bu süreç, modelin doğruluğunu artırmak için tekrarlanabilir.

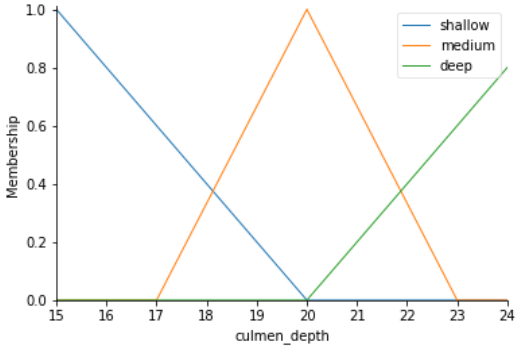
Doğruluk hesaplama adımı, modelin ne kadar başarılı olduğunu değerlendirir. Elde edilen sonuçlar, modelin güvenilirliği hakkında bilgi sağlar.

Son olarak, grafikler oluşturulur. Bu grafikler, modelin doğruluğu, tahmin edilen cinsiyet dağılımı, gerçek cinsiyet dağılımı gibi önemli görsel bilgileri sunar ve analiz edilmesini sağlar.

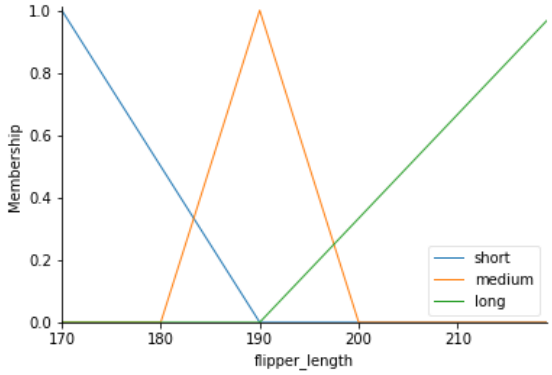
Bu süreç, penguen cinsiyeti tahmininde kullanılan bulanık mantık modelinin nasıl çalıştığını ve performansının nasıl değerlendirildiğini ayrıntılı bir şekilde açıklar.



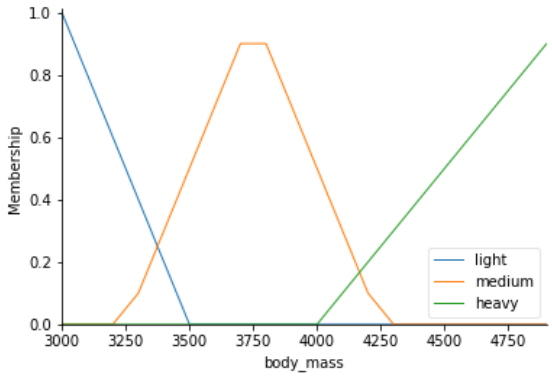
Şekil 5 *Gaga Uzunluk Bulanık Mantık Modeli*



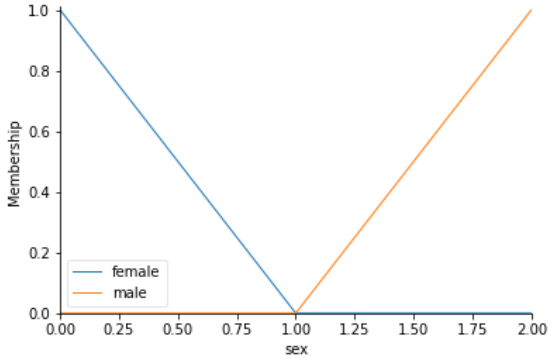
Şekil 6 *Gaga Derinliği Bulanık Mantık Modeli*



Şekil 7 *Kanat Uzunluk Bulanık Mantık Modeli*



Şekil 8 *Vücut Ağırlık Bulanık Mantık Modeli*



Şekil 9 *Cinsiyet Bulanık Mantık Modeli*

MLPClassifier sınıfı, çok katmanlı perceptron (MLP) sınıflandırıcısını oluşturur. Bu sınıf, gizli katmanların boyutunu, maksimum iterasyon sayısını ve başlangıç ağırlıklarını belirlemek için kullanılır. fit yöntemi, bu modeli eğitmek için kullanılır. Eğitim verileri (x_train ve y_train) bu yönteme iletilir ve model, bu verilere göre ayarlanır. Model eğitildikten sonra, predict yöntemi, modelin test seti üzerindeki performansını değerlendirmek için kullanılır. Test seti verileri (x_test) bu yönteme iletilir ve model, bu verilere dayanarak tahminler yapar. Bu tahminler daha sonra gerçek sınıflarla (y_test) karşılaştırılarak doğruluk skoru hesaplanır.

Son olarak, accuracy_score fonksiyonu, modelin doğruluğunu değerlendirmek için kullanılır. Bu fonksiyon, gerçek sınıflarla modelin tahmin ettiği sınıfları karşılaştırır ve doğruluk skorunu hesaplar. Bu, modelin ne kadar doğru tahminler yaptığını ölçer ve modelin performansını değerlendirmek için yaygın bir metrik olarak kullanılır.

Bu fonksiyonlar, makine öğrenimi modelinin eğitiminden başlayarak doğruluğunun değerlendirilmesine kadar olan kritik adımları gerçekleştirmek için bir araçlar sağlar. Bu adımlar, modelin başarısını değerlendirmek ve iyileştirmek için önemlidir, bu da daha güvenilir ve etkili sonuçlar elde etmeyi sağlar.

Algoritmalar	Doğruluk
LVQ	0.90662
ANFIS	0.68373
ANN(MLP)	0.89552

init	time	inertia	homo	compl	v-meas	ARI	AMI	silhouette
k-means++	0.038s	1169069.5230511124	0.737	0.745	0.741	0.657	0.739	0.176
random	0.016s	1165461.2695883983	0.733	0.740	0.737	0.659	0.734	0.185
PCA-based	0.023s	1188310.4292529237	0.784	0.787	0.785	0.741	0.783	0.183

Kaynakça

[1]Woods Hole Oceanographic Institution. (n.d.). Penguin Life Observatories. Retrieved from https://www2.whoi.edu/site/mars/penguins/

[2]Viblanc, Vincent A., et al. "Body girth as an alternative to body mass for establishing condition indexes in field studies: a validation in the king penguin." Physiological and Biochemical Zoology 85.5 (2012): 533-542.

[3] Wallace, Roberta S., et al. "Morphometric determination of gender in adult Humboldt Penguins (Spheniscus humboldti)." Waterbirds (2008): 448-453.

[4] Le, Hieu M., et al. "Weakly Labeling the Antarctic: The Penguin Colony Case." CVPR Workshops. 2019.

[5] Zhang J, O'Reilly KM, Perry GLW, Taylor GA, Dennis TE (2015) Extending the Functionality of Behavioural Change-Point Analysis with k-Means Clustering: A Case Study with the Little Penguin (Eudyptula minor). PLoS ONE 10(4): e0122811. https://doi.org/10.1371/journal.pone.0122811

[6]Clucas, Gemma V., et al. "Comparative population genomics reveals key barriers to dispersal in Southern Ocean penguins." Molecular Ecology 27.23 (2018): 4680-4697.

[7]Younger, Jane L., et al. "The challenges of detecting subtle population structure and its importance for the conservation of emperor penguins." Molecular Ecology 26.15 (2017): 3883-3897.

[8]Jadwiszczak, Piotr, and Carolina Ileana Alicia Acosta Hospitaleche. "Distinguishing between two Antarctic species of Eocene Palaeeudyptes penguins: a statistical approach using tarsometatarsi." Polish Polar Research 34 (2013).

[9]Schreer, Jason F., et al. "Classification of Dive Profiles: A Comparison of Statistical Clustering Techniques and Unsupervised Artifical Neural Networks." *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 3, no. 4, 1998, pp. 383–404. *JSTOR*, https://doi.org/10.2307/1400572. Accessed 8 May 2024.

[10] Hart, T., Mann, R., Coulson, T. et al. Behavioural switching in a central place forager: patterns of diving behaviour in the macaroni penguin (Eudyptes chrysolophus). Mar Biol 157, 1543–1553 (2010). https://doi.org/10.1007/s00227-010-1428-2

[11] Chessa, S., Micheli, A., Pucci, R., Hunter, J., Carroll, G., & Harcourt, R. (2017). A comparative analysis of SVM and IDNN for identifying penguin activities. Applied Artificial Intelligence, 31(5-6), 453-471.

[12] Le, H. M., Gonçalves, B. C., Samaras, D., & Lynch, H. J. (2019, June). Weakly Labeling the Antarctic: The Penguin Colony Case. In CVPR Workshops (pp. 18-25).

[13] Youssef Aboelwafa."Clustering Penguin Species". Kaggle. [Online]. URL: https://www.kaggle.com/datasets/youssefaboelwafa/clustering-penguins-species

[14] Celebi, M. Emre, Hassan A. Kingravi, and Patricio A. Vela. "A comparative study of efficient initialization methods for the k-means clustering algorithm." Expert systems with applications 40.1 (2013): 200-210.

[15]Pham, Duc Truong, Stefan S. Dimov, and Chi D. Nguyen. "Selection of K in K-means clustering." Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science 219.1 (2005): 103-119.

[16] T. Kohonen, "Improved versions of learning vector quantization," 1990 IJCNN International Joint Conference on Neural Networks, San Diego, CA, USA, 1990, pp. 545-550 vol.1, doi: 10.1109/IJCNN.1990.137622.

[17] Nova, David, and Pablo A. Estévez. "A review of learning vector quantization classifiers." Neural Computing and Applications 25 (2014): 511-524.

[18] M. A. Denai, F. Palis and A. Zeghbib, "ANFIS based modelling and control of non-linear systems : a tutorial," 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583), The Hague, Netherlands, 2004, pp. 3433-3438 vol.4, doi: 10.1109/ICSMC.2004.1400873.

[19] M. Alizadeh, M. Lewis, M. H. F. Zarandi and F. Jolai, "Determining significant parameters in the design of ANFIS," 2011 Annual Meeting of the North American Fuzzy Information Processing Society, El Paso, TX, USA, 2011, pp. 1-6, doi: 10.1109/NAFIPS.2011.5751958.

[20] Malik, Nitin. "Artificial neural networks and their applications." arXiv preprint cs/0505019 (2005).

[21] Yang, Guangyu Robert, and Xiao-Jing Wang. "Artificial neural networks for neuroscientists: a primer." Neuron 107.6 (2020): 1048-1070.