Python ile Yapay Zekâ Uygulamaları Dersi Proje Raporu

2022-2023 Güz Yarıyılı

Ad: Emre

Soyad: Şimşek

Numara: 23360859806

1.Özet

Bu çalışma, "bodyPerformance" adlı veri setini kullanarak bireylerin çeşitli özelliklerini değerlendirerek belirli vücut sınıflarını tespit etmeyi amaçlamaktadır. Üç farklı vücut sınıfı tipi, sırasıyla 0 (iyi), 1 (orta) ve 2 (kötü) olarak tanımlanmıştır.

Sınıflandırma işleminde Rastgele Orman, Lojistik Regresyon ve Destek Vektör Makineleri algoritmaları kullanılmıştır. Bu algoritmaların seçilme sebebi, literatürdeki diğer algoritmalarla karşılaştırıldığında yüksek doğruluk oranlarına sahip olmalarıdır.

Random Forest, bir ensemble (topluluk) öğrenme algoritmasıdır. Birden fazla karar ağacını birleştirerek daha güçlü ve genelleştirilebilir bir model oluşturur. Her bir ağaç, rastgele örneklemeler ve rastgele özellik seçimleri kullanarak eğitilir. Bu sayede, overfitting (aşırı öğrenme) riski azalır ve genel performans artar.

Lojistik Regresyon, sınıflandırma problemleri için kullanılan bir istatistiksel modeldir. Giriş verilerini bir veya birden fazla bağımsız değişkenle ilişkilendirir ve bu ilişkiyi kullanarak bir çıktı elde eder.

Destek Vektör Makineleri, sınıflandırma ve regresyon problemleri için kullanılan bir öğrenme algoritmasıdır. Temelde, veri noktalarını sınıflandırmak için en iyi ayrım hattını (veya hiper düzlemi) bulmaya çalışır. Çok boyutlu veri setlerinde etkilidir ve özellikle yüksek boyutlu uzaylarda iyi performans gösterir.

Bu çalışma, "bodyPerformance" veri seti üzerinde Rastgele Orman, Lojistik Regresyon ve Destek Vektör Makineleri algoritmalarını kullanarak bir sınıflandırma gerçekleştirmiştir. Yapılan sınıflandırma, sağlık uzmanları, spor koçları ve bireylerin kişisel sağlıklarını izlemek istedikleri durumlarda etkili rehberlik sağlama konusunda yardımcı olmaktadır.

Bu algoritmaların uygulamalarındaki başarı, genel olarak veri setindeki bireylerin vücut sınıflarını doğru bir şekilde belirleme yeteneklerini yansıtmaktadır. %70 üzerinde elde edilen başarı oranları, bu algoritmaların sağlık uzmanlarının bireylerin sağlık durumlarını değerlendirmelerine, spor koçlarının kişiselleştirilmiş antrenman programları oluşturmalarına ve bireylerin kendi sağlıklarını izlemelerine yardımcı olma konusundaki potansiyelini vurgulamaktadır.

2.Veriler

Veri setimiz, bireylerin çeşitli fiziksel özelliklerini ve performans metriklerini içeren bir tabloyu temsil eder. Toplamda 13,393 örnekten oluşmaktadır. Her bir örnek, bireyin yaşını, cinsiyetini, boyunu, kilosunu, vücut yağ yüzdesini, kan basıncını, kavrama gücünü, esneklik yeteneklerini, şınav sayısını ve geniş atlayış mesafesini içerir. Ayrıca, her bir bireyin genel bir performans sınıflandırmasını ifade eden bir "class" sütunu da bulunmaktadır.

Veri setinin detaylı olarak sütunları şu şekildedir:

1. age: Bireyin yaşını temsil eden sayısal değerler.
2. gender: Bireyin cinsiyetini temsil eden "M" (erkek) veya "F" (kadın) değerleri.
3. height_cm: Bireyin boyunu temsil eden sayısal değerler (santimetre cinsinden).
4. weight_kg: Bireyin kilosunu temsil eden sayısal değerler (kilogram cinsinden).
5. body fat_%: Bireyin vücut yağ yüzdesini temsil eden sayısal değerler.
6. diastolic: Bireyin diyastolik kan basıncını temsil eden sayısal değerler.
7. systolic: Bireyin sistolik kan basıncını temsil eden sayısal değerler.
8. gripForce: Bireyin kavrama gücünü temsil eden sayısal değerler.
9. sit and bend forward_cm: Bireyin oturup öne eğilme mesafesini temsil eden sayısal değerler (santimetre cinsinden).
10. sit-ups counts: Bireyin şınav sayısını temsil eden sayısal değerler.
11. broad jump_cm: Bireyin geniş atlayış mesafesini temsil eden sayısal değerler (santimetre cinsinden).
12. class: Bireyin genel performans sınıflandırmasını temsil eden harf (A, B, C, D) değerleri.

Bu projede, veri setinde yer alan dört farklı sınıf (A, B, C, D) içerisinde, model performansının geliştirilmesi amacıyla bir düzenleme yapılmıştır. Bu düzenleme, özellikle B ve C sınıflarının birleştirilip "orta sınıf" olarak düşünülebilecek bir sınıf oluşturulmuştur. Bu yaklaşım, sınıf etiketlerinin daha önceki sınıflandırmadan farklı bir yapıda ele alınmasını sağlayarak, modelin daha etkili bir şekilde öğrenme yapmasına ve genelleme yeteneğini artırmasına yöneliktir. Bu sınıf birleştirme stratejisi ile modelin performansını geliştirmeyi amaçlamaktadır.

3.Yöntem

Çalışmada ele alınan problem tipi, veri setindeki bireyleri belirli sınıflara (A, (B-C), D) sınıflandırma görevidir. Temel amaç, bireylerin genel fiziksel durumlarına dayalı olarak uygun sınıflara atanmalarını sağlamaktır. Bu, özellikle sağlık ve spor performansıyla ilgili değerlendirmelerde kullanılabilecek bir sınıflandırma problemini içermektedir.

Kullandığım Algoritmalar

Random Forest, makine öğrenimi alanında sıkça kullanılan bir ensemble öğrenme algoritmasıdır. Birden fazla karar ağacını bir araya getirerek daha güçlü ve genelleştirilmiş bir model oluşturmayı amaçlar. Her bir karar ağacı, rastgele seçilen örnekler ve özellikler üzerinde eğitilir, bu da overfitting (aşırı öğrenme) riskini azaltır. Sonuç olarak, Random Forest algoritması, veri setindeki karmaşıklığı azaltarak yüksek doğrulukta tahminler yapabilen bir model oluşturur. Hem sınıflandırma hem de regresyon problemlerinde etkili bir şekilde kullanılabilir.

Random Forest, karar ağaçlarının zayıflıklarını gidererek daha güçlü bir model oluşturur ve genelleme performansını artırır. Özellikle, özellik seçiminde etkili olup önemli özellikleri belirlemede kullanılabilir. Ayrıca, veri setindeki aykırı değerlere karşı dayanıklıdır. Eğitim hızı diğer ensemble öğrenme algoritmalarına göre daha hızlıdır ve parametre ayarı diğer algoritmalara göre daha kolaydır. Bu avantajları, yüksek boyutlu veri setlerinde bile etkili bir şekilde kullanılmasını sağlar.

Bu özellikleri sayesinde Random Forest, birçok alanda başarıyla kullanılmaktadır. Örneğin, sağlık sektöründe hastalık teşhisi, finans sektöründe dolandırıcılık tespiti, perakende sektöründe müşteri davranışlarının analizi ve üretim sektöründe kalite kontrolü gibi alanlarda yaygın olarak tercih edilmektedir.

Lojistik regresyon, istatistik ve makine öğrenimi disiplinlerinde geniş bir kullanıma sahip olan bir modelleme tekniğidir. Temelde, bağımsız değişkenlerin kombinasyonlarına dayalı olarak belirli bir olayın gerçekleşme olasılığını tahmin etmek amacıyla kullanılır. Bu algoritma genellikle binary classification (iki sınıf arasındaki ayrım) problemlerinde tercih edilirken, aynı zamanda regresyon problemlerini çözmek için de adapte edilebilmektedir.

Lojistik regresyon, sınıflandırma problemlerinde kullanılan temel matematiksel ifade olan lojistik fonksiyonu (sigmoid fonksiyonu) içermektedir. Sigmoid fonksiyonu, bir giriş değerini alır ve onu 0 ile 1 arasında bir olasılık değerine dönüştürerek sınıflandırma gerçekleştirir. Bu, binary classification problemlerinde, bir olayın gerçekleşme olasılığını belirlemek için kullanılır.

Lojistik regresyon, sadece sınıflandırma ile sınırlı kalmayıp aynı zamanda regresyon problemlerini çözmek üzere de kullanılabilen esnek bir modelleme yaklaşımına sahiptir. Regresyon problemlerinde, bağımlı değişkenin sürekli bir sayısal değeri tahmin etmek amacıyla kullanılır.

Bu modelin avantajları arasında yüksek yorumlanabilirlik, hızlı eğitim süreçleri ve düşük hesaplama maliyetleri bulunmaktadır. Ancak, doğrusal sınırlamaları nedeniyle karmaşık ilişkileri modelleme konusunda bazı sınırlamalara sahiptir. Lojistik regresyon, sınıflandırma ve regresyon problemlerine etkili bir çözüm sunmasıyla, istatistiksel analiz ve makine öğrenimi uygulamalarında sıkça kullanılan bir araç haline gelmiştir.

Destek Vektör Makineleri (SVM), makine öğrenimi alanında yaygın olarak kullanılan bir sınıflandırma algoritmasıdır. Temel olarak, veri noktalarını sınıflara ayırmak için bir karar sınırı belirleme amacını taşır ve bu karar sınırını oluştururken destek vektörlerini kullanır.

SVM'nin temel amacı, veri noktalarını iki veya daha fazla sınıfa ayırmak ve bu sınıfları birbirinden net bir şekilde ayıran bir karar sınırı oluşturmaktır. Bu süreç, veri noktalarının bir uzayda konumlandırılması ve ardından optimal bir karar sınırının belirlenmesiyle gerçekleşir. Bu sınır, veri noktalarını sınıflandırmada maksimum ayrımı sağlamak üzere tasarlanmıştır.

Destek vektörleri, karar sınırına en yakın noktalardır ve genellikle bu noktalar arasındaki mesafeye "marj" denir. SVM, bu marjı maksimize ederek, modelin genelleme yeteneğini artırır ve aşırı öğrenmeye karşı direnç kazandırır.

Öne çıkan bir özellik, SVM'nin doğrusal ve doğrusal olmayan veri setleri üzerinde etkili bir şekilde çalışabilmesidir. Doğrusal SVM, veri noktalarını doğrusal bir sınır kullanarak sınıflandırırken, doğrusal olmayan SVM, çekirdek fonksiyonları aracılığıyla daha karmaşık veri yapılarını modellemek için kullanılır.

SVM'nin başarıları, özellikle yüksek boyutlu veri setlerinde ve gürültülü veri ortamlarında, diğer sınıflandırma yöntemlerine göre üstünlüğünü göstermiştir. Ancak, büyük veri setlerinde eğitim süresinin uzunluğu ve hiperparametre ayarının hassasiyeti gibi zorluklar da vardır.

Sonuç olarak, SVM, sınıflandırma problemlerini çözmek için güçlü ve esnek bir araç olarak öne çıkmaktadır. Ancak, spesifik uygulama bağlamında dikkatlice ayarlanması ve kullanılması gereken bir algoritmadır.

4.Uygulama

Veri analitiği sürecinde, gender ve class sütunlarındaki sayısal olmayan değerleri uygun bir formata dönüştürmek amacıyla LabelEncoder kütüphanesini kullanarak bir ön işleme gerçekleştirdik. Bu sayede, veri setimizdeki kategorik değerleri numerik değerlere dönüştürerek, makine öğrenimi algoritmalarının daha etkili bir şekilde çalışmasını sağladık.

Eğitim ve test verilerini ayırmak için genellikle veri setimizi %80 eğitim ve %20 test oranında böleriz. Bu oranlar, modelin öğrenme yeteneğini optimize etmede genellikle iyi bir denge sağlar.

Eksik verilerin varlığını belirlemek amacıyla veri setini incelendiğinde herhangi bir eksik veri tespit edilmedi. Bu, modelin eğitim ve test aşamalarında veri bütünlüğünün sağlanmasına yönelik güvenilir bir temel oluşturdu. Eksiksiz bir veri seti, makine öğrenimi algoritmalarının daha tutarlı ve güvenilir sonuçlar üretmesine olanak tanır.

Hiper parametre seçimi, modelin performansını belirleyen önemli bir faktördür. Bu aşamada, veri setine uygun olarak öğrenme oranı ve diğer hiper parametrelerin belirlenmesi için çeşitli denemeler gerçekleştirildi. Her bir hiper parametre seçimi, modelin genel başarı düzeyini optimize etme amacı güderek, eğitim ve test verileri üzerindeki performansını artırmaya yönelik bir strateji izlendi. Bu süreç, modelin doğruluğunu ve genelleme yeteneğini artırmak amacıyla deneme-yanılma yöntemini içeriyordu.

Lojistik regresyon için seçilen ve uygulanan hiper parametreler şunlardır:

1. **solver:** Optimizasyon problemi çözülürken kullanılacak algoritmayı belirten bir parametredir.
2. **multi_class:** Çoklu sınıflı (multiclass) problemlerde kullanılacak stratejiyi belirten bir parametredir.

3. **max_iter:** Optimizasyon algoritmasının maksimum iterasyon sayısını belirten bir parametredir.

Destek Vektör Makineleri için seçilen ve uygulanan hiper parametreler şunlardır:

1. **C:** SVM modelinin düzenlileştirmesini kontrol eden bir hiper parametredir.
2. **kernel:** SVM'nin giriş verilerini daha yüksek boyutlu uzaylara taşıyan bir matematiksel işlemdir. Bu, doğrusal olarak ayrılamayan veri setlerini ele almak için kullanılır.
3. **degree:** Polinom çekirdeği seçildiğinde kullanılan polinomun derecesini belirten bir parametredir.

Rassal Orman için seçilen ve uygulanan hiper parametreler şunlardır:

1. **n_estimators:** Orman içindeki ağaç sayısını belirler.
2. **max_depth:** Her bir karar ağacının maksimum derinliğini sınırlayan bir parametredir.
3. **min_samples_split:** Bir düğümün iki alt düğüme ayrılabilmesi için gereken minimum örnek sayısını belirler.
4. **min_samples_leaf:** Bir yaprağın minimum örnek sayısını belirler. Yaprak, bir ağacın en altındaki düğümlerdir.

Bu şekilde gerçekleştirilen veri işleme ve modelleme aşamaları, hem veri setinin uygun bir şekilde hazırlanmasını sağlamış hem de kullanılan makine öğrenimi modelinin daha sağlıklı ve güvenilir sonuçlar üretmesine olanak tanımıştır.

5.Sonuçlar

Destek Vektör Makineleri algoritması sonuçları 2 farklı model üzerinde K-Fold Cross-Validation kullanılarak şu şekilde elde edilmiştir:

Ortalama Eğitim verisi doğruluk oranı: 0.74
Ortalama KFold doğruluk oranı: 0.73

Lojistik Regresyon algoritmasında sonuçlar şu şekildedir:

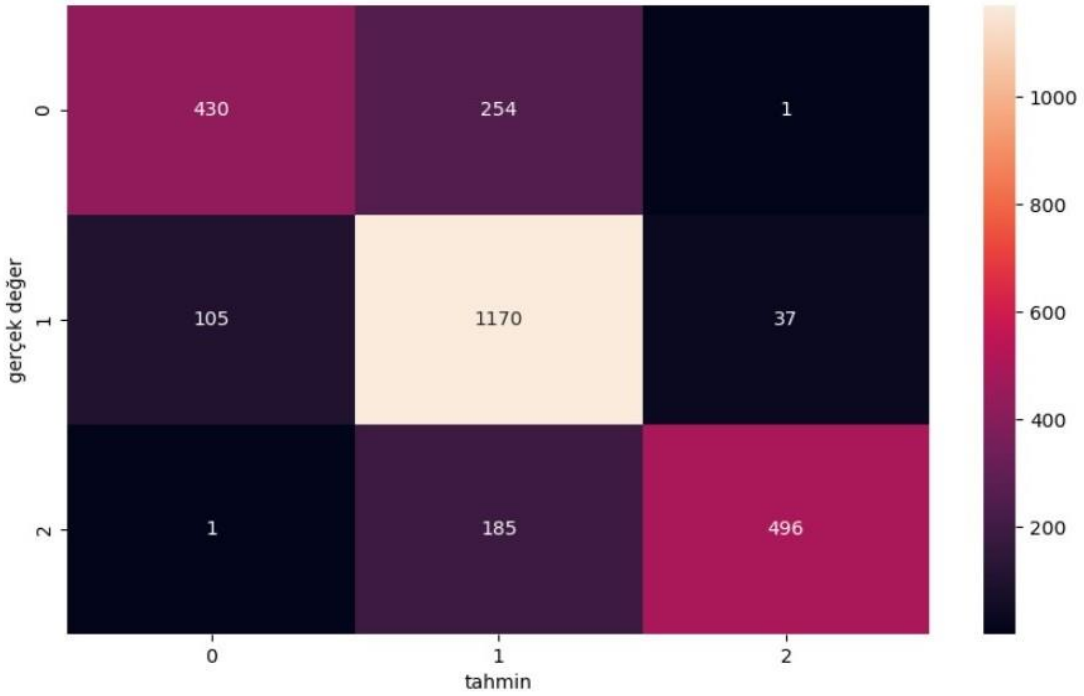
Ortalama Eğitim verisi doğruluk oranı: 0.729
Ortalama KFold doğruluk oranı: 0.725
Bu durum,2 farklı algoritmanın eğitim sırasında öğrenilen bilgilerin test verisinde de genel bir doğrulukla uyuştuğunu göstermektedir.

Rassal Orman modelinde sonuçlar şu şekildedir:

Ortalama Eğitim verisi doğruluk oranı: 0.951
Ortalama KFold doğruluk oranı: 0.808

Eğitim verilerindeki yüksek doğruluk oranına karşılık, test verilerindeki doğruluk oranı daha düşük. Bu durum, modelin eğitim verilerine aşırı uyum sağlayarak genelleme yeteneğini kaybettiğini ve yeni, görülmemiş verilere karşı daha zayıf olduğunu gösterir (overfitting).

Bu sorun hiper parametre ayarlaması yapılarak farklı bir model üzerinde düzeltilmiştir. Sonuçlar şu şekildedir:
Rassal Orman Eğitim verisi doğruluk oranı: 0.808
Rassal Orman Test verisi doğruluk oranı: 0.809



Yukarıdaki tabloda test verileri için doğru olan değerler ve tahmin değerleri gösterilmiştir. Ana köşegen üzerinde bulunan değerler doğru tahminlerin bulunduğu kısımdır diğer kısımlarda bulunan değerler yanlış tahminleri içermektedir. Rassal Orman algoritmamız, veri setindeki örneklerin yaklaşık olarak %81ini doğru bir şekilde sınıflandırarak başarılı sonuçlar elde etti. Bu oran, modelin genel performansının kabul edilebilir düzeyde olduğunu gösteriyor ancak, modelin daha da iyileştirilebilmesi için hiper parametre ayarlamaları veya veri setinde farklı teknikler değerlendirilebilir.