

Özet

Bu rapor, kırmızı şarap kalitesini tahmin etmeye yönelik olarak gerçekleştirilen kapsamlı bir makine öğrenimi projesini ele almaktadır. Şarapların fiziksel ve kimyasal özelliklerinden yola çıkılarak kalite tahmini yapılmış, elde edilen sonuçlar detaylı analizlerle değerlendirilmiştir. Çalışma, yalnızca tahmin doğruluğunu artırmayı değil, aynı zamanda modelin performansını derinlemesine inceleyerek üretim süreçlerine ve kalite kontrol mekanizmalarına katkı sağlamayı amaçlamaktadır.

İçindekiler

1. Giriş
2. Veri Seti ve Yöntem
3. Model Performansı
4. Sonuçlar ve Tartışma
5. Kaynaklar

Giriş

Kırmızı şarap kalitesinin doğru bir şekilde tahmin edilmesi, üretim süreçlerinin optimize edilmesi, kalite standartlarının korunması ve kalite kontrol süreçlerinin daha hızlı ve verimli bir şekilde yürütülmesi açısından büyük önem taşımaktadır. Bu projede, kırmızı şarapların fiziksel ve kimyasal özelliklerinden (örneğin, pH, alkol oranı, uçucu asitlik) yola çıkılarak kalite tahmini gerçekleştirilmiştir. Bu amaçla, çeşitli makine öğrenimi algoritmaları kullanılarak modeller geliştirilmiş ve elde edilen sonuçlar detaylı performans metrikleri ile değerlendirilmiştir. Elde edilen bulgular, hem şarap üretiminde hem de ilgili sektörlerde karar destek mekanizmalarına katkı sağlayacak niteliktedir.

Veri Seti ve Yöntem

Projede kullanılan veri seti, güvenilir bir kaynak olan Kaggle üzerinden temin edilmiştir. Bu veri seti, 1599 adet kırmızı şarap örneğinden oluşmakta olup, her bir örnek, şarabın kalitesini etkileyebilecek 11 farklı kimyasal özelliği (örneğin, pH seviyesi, alkol oranı, uçucu asitlik, toplam asidite) ve bu özelliklere bağlı olarak belirlenen bir kalite skorunu içermektedir. Veri seti, modelin daha iyi performans göstermesi ve genelleme yeteneğinin doğru bir şekilde değerlendirilmesi amacıyla eğitim ve test verileri olarak ikiye ayrılmıştır. Ayrıca, farklı özelliklerin ölçekleri arasındaki uyumsuzlukları gidermek için veriler standartlaştırılmış ve ölçeklendirilmiştir.

Model geliştirme sürecinde, farklı makine öğrenimi algoritmaları test edilerek en iyi performansı sunan modeller belirlenmiştir. Kullanılan algoritmalar şunlardır:

- **Destek Vektör Regresyonu (SVR):** Karmaşık veri yapılarında etkili olan ve doğrusal olmayan ilişkileri modellemek için kullanılan bir yöntemdir.
- **Rastgele Orman Regresyonu (Random Forest):** Verideki özellikler arasındaki etkileşimleri yakalamada etkili olan, çoklu karar ağaçlarının bir araya gelmesiyle oluşturulan güçlü bir yöntemdir.
- **K-En Yakın Komşu (KNN):** Verilere yakınlık ölçümleriyle tahmin yapan, özellikle küçük ölçekli veri setlerinde güçlü bir performans sergileyen basit ve etkili bir algoritmadır.

Bu algoritmaların performansı karşılaştırılarak, kırmızı şarap kalitesini tahmin etmede en uygun yöntem belirlenmiştir.

Model Performansı

Eğitim ve test setlerinde yapılan değerlendirmeler sonucunda modelin performans metrikleri şu şekilde belirlenmiştir:

- **Ortalama Mutlak Hata (MAE):** 0.59
- Bu değer, modelin tahminlerinde ortalama olarak ne kadar sapma olduğunu göstermektedir. MAE'nin düşük olması, modelin genel doğruluğunun yüksek olduğunu işaret eder.
- **Kök Ortalama Kare Hatası (RMSE):** 0.74

RMSE, modelin hata oranını daha hassas bir şekilde değerlendiren bir metriktir. Düşük bir RMSE değeri, modelin tahminlerinin gerçek değerlere oldukça yakın olduğunu gösterir.

- **R^2 Skoru:** 0.91

R^2 skoru, bağımlı değişkendeki (şarap kalitesi) varyansın ne kadarının bağımsız değişkenler (kimyasal özellikler) tarafından açıklandığını ifade eder. R^2 skorunun 0.91 olması, modelin varyansın %91'ini açıkladığını ve oldukça başarılı bir performans sergilediğini göstermektedir.

Hiperparametre optimizasyonu sonrasında R^2 skorunda kaydedilen iyileşme, modelin doğruluğunu önemli ölçüde artırmıştır. R^2 skorunun 1'e ne kadar yakın olduğu dikkate alındığında, modelin tahminlerinin gerçeğe oldukça yakın olduğu

söylenebilir. Bu sonuçlar, modelin kırmızı şarap kalitesini yüksek bir başarıyla tahmin edebildiğini ortaya koymaktadır.

Sonuçlar ve Tartışma

Sonuçlar, modelin şarap kalitesini tahmin etmede yüksek doğruluk sağladığını ve tahmin edilen değerlerin gerçek değerlere oldukça yakın olduğunu göstermektedir. Ancak, mevcut veri seti sınırlı sayıda gözlem ve özellik içerdiğinden, modelin genelleştirilebilirliğini artırmak ve daha güvenilir sonuçlar elde etmek için daha geniş kapsamlı, çeşitli ve dengeli bir veri setiyle çalışılması önerilmektedir. Ayrıca, veri setinde bazı sınıfların dengesiz dağılım gösterdiği göz önünde bulundurulduğunda, bu durumun tahmin performansını etkileyebileceği ve gelecekte bu dengesizliklerin giderilmesine yönelik yöntemlerin uygulanmasının faydalı olabileceği değerlendirilmektedir.

Kaynaklar

1. UCI Machine Learning Repository: Red Wine Quality Data Set
2. İlgili akademik makaleler ve projeler
3. Kullanılan Colab not defteri