MNIST Veri Seti ile Sayı Tanıma Projesi: Kirlilik Analizi ve En İyi Model Belirleme

Giriş

Dijital çağın hızla ilerlemesiyle birlikte, otomatik sayı tanıma sistemleri, birçok önemli uygulama alanında kritik bir rol oynamaktadır. Özellikle, el yazısı rakamların tanınması, banka çeklerinin işlenmesinden, posta kodlarının sınıflandırılmasına kadar birçok alanda bu sistemler kullanılmaktadır. Bu sistemler, insan hatalarını azaltarak iş verimliliğini artırırken, aynı zamanda hız, doğruluk ve güvenilirlik açısından da büyük avantajlar sunmaktadır.

Bu bağlamda, sayısal verilerin tanınması ve sınıflandırılması için en çok kullanılan veri setlerinden biri olan MNIST (Modified National Institute of Standards and Technology) veri seti, araştırmacılar ve veri bilimciler için standart bir benchmark olarak hizmet etmektedir. MNIST veri seti, el yazısı rakamların (0-9 arası) 70,000 görüntüsünden oluşur ve bu görüntüler 28x28 piksel boyutundadır. Bu veri seti, sayı tanıma algoritmalarının eğitilmesi ve değerlendirilmesi için geniş bir kullanım alanına sahiptir.

MNIST veri seti üzerinde yapılan çalışmalar, sayı tanıma sistemlerinin geliştirilmesinde önemli bir rol oynamaktadır. Bu çalışmalar, yeni algoritmaların geliştirilmesi, mevcut algoritmaların karşılaştırılması ve iyileştirilmesi için değerli bir kaynak sağlamaktadır. Ayrıca, MNIST veri seti, öğrenme algoritmalarının genelleme yeteneğini ve sağlamlığını test etmek için de sıkça kullanılmaktadır.

Algoritmalar ve Performans Değerlendirme

Bu proje kapsamında, dört farklı makine öğrenimi algoritması (KNN, SVC, MLPC ve Decision Tree Classifier) temiz ve kirli veri setleri üzerinde eğitilmiş ve performansları kapsamlı bir şekilde değerlendirilmiştir. Her bir modelin performansı, doğruluk (accuracy) metriği kullanılarak değerlendirilmiştir. Ayrıca, her bir model için hiperparametrelerin optimize edilmesi ve çapraz doğrulama (cross-validation) gibi tekniklerin kullanılmasıyla daha sağlam sonuçlar elde edilmiştir. Temiz veri seti üzerinde yapılan deneyler, her bir modelin başlangıç performansını belirlemek için gerçekleştirilmiştir. Daha sonra, veri setine eklenen rastgele gürültülerle oluşturulan kirli veri seti üzerinde modellerin performansı test edilmiştir.

En İyi Model Belirleme

Kirli veri seti üzerinde yapılan deneyler sonucunda, modellerin performansları karşılaştırılmış ve en iyi performansı gösteren model belirlenmiştir. Bu karar, modellerin kirli veri setindeki doğruluklarına dayanarak alınmıştır. Ancak, sadece doğruluk metriği kullanılmamış; aynı zamanda modelin kararlılığı, hiperparametrelerin ayarlanabilirliği ve hesaplama süreleri gibi faktörler de göz önünde bulundurulmuştur. En iyi model, kirli veri setinde en yüksek doğruluk oranına sahip olan model olarak tanımlanmıştır.

Sonuçlar

Kirli veri seti üzerinde yapılan performans değerlendirmesinde, SVC modelinin en iyi performansı gösterdiği belirlenmiştir. Diğer modellerin performansı da incelenmiş olmasına rağmen, SVC'nin kirli veri setindeki başarımının diğerlerine göre belirgin şekilde daha yüksek olduğu tespit edilmiştir. Ancak, bu sonuçların genelleme yeteneği, hiperparametrelerin optimize edilmesi, ve modelin hesaplama süreleri gibi faktörlerle birlikte değerlendirilmesi gerekmektedir.

Tartışma

SVC'nin kirli veri setindeki başarısı, modelin karmaşıklığı ve esnekliği ile ilişkilendirilebilir. Özellikle, doğrusal olmayan karar sınırları oluşturabilen RBF çekirdek fonksiyonu kullanılarak uygulandığında, SVC'nin çeşitli görüntü bozulmalarını ve gürültüleri etkili bir şekilde sınıflandırabileceği düşünülmektedir. Bu durum, modelin kirli veri setindeki performansını artırarak, gerçek dünya verileriyle karşılaşıldığında daha sağlam sonuçlar üretebilme yeteneğini göstermektedir.

Sonuç

Bu çalışma, MNIST veri setine eklenen kirli veri seti üzerinde dört farklı makine öğrenimi modelinin performansını değerlendirmiştir. Kirli veri seti üzerinde yapılan deneyler sonucunda, SVC modelinin en iyi performansı gösterdiği belirlenmiştir. Bu sonuçlar, gerçek dünya verileriyle model performansını değerlendirirken, modelin sağlamlığı ve genelleme yeteneği gibi faktörlerin önemini vurgulamaktadır. SVC modelinin başarısı, özellikle el yazısı rakamların tanınması gibi karmaşık ve gürültülü veri setleri üzerinde de etkili bir şekilde çalışabileceğini göstermektedir. Bu bulgular, otomatik sayı tanıma sistemlerinin güvenilirliğini artırmak için daha karmaşık ve esnek modellerin kullanılmasının önemini vurgular niteliktedir.