İyi Bir Veri Seti Nasıl Hazırlanır?

Bir yapay zekâ modelinin başarısını en çok belirleyen unsurlardan biri veri setinin kalitesidir. Modelin doğru öğrenebilmesi için veri seti yalnızca "çok sayıda görsel" içermemeli; aynı zamanda temiz, dengeli, tutarlı ve doğru etiketlenmiş olmalıdır. Bu nedenle veri seti hazırlama süreci; veri toplama, temizleme, standartlaştırma ve etiketleme adımlarından oluşan sistemli bir çalışmadır.

İlk adım problem tanımını netleştirmektir. Modelin amacı belirlenmelidir: sınıflandırma mı yapılacak, nesne tespiti mi, yoksa segmentasyon mu? Örneğin sınıflandırmada tek etiket yeterliyken, multi-label senaryolarda bir görsel birden fazla etiket alabilir (class=hoodie, color=black gibi). Segmentasyon için ise maske formatı ve sınıf tanımları baştan netleştirilmelidir.

Veri toplama aşamasında çeşitlilik kritik öneme sahiptir. Görseller, modelin gerçek hayatta karşılaşacağı koşulları temsil etmelidir. Bu nedenle farklı ışık ortamları, arka planlar, açı ve uzaklıklar içeren örnekler seçilmelidir. Veri çeşitliliği düşük olursa model ezber yapabilir ve yeni görsellerde başarısız olur.

Toplanan veriler mutlaka **kalite kontrolünden** geçirilmelidir. Çok düşük çözünürlüklü, aşırı karanlık veya aşırı parlak (pozlama hatalı), bulanık ya da bozuk görseller modelin öğrenmesini olumsuz etkiler. Bu tip veriler mümkünse elenmeli veya iyileştirilmelidir. Ayrıca veri setinde tekrar eden (duplicate) görseller bulunmamalıdır. Aynı görselin tekrar edilmesi veri setini şişirir, veri çeşitliliğini düşürür ve değerlendirme sonuçlarını yanıltabilir.

Etiketleme aşaması veri setinin en kritik kısmıdır. Etiketlerin doğru ve tutarlı olması gerekir. Aynı sınıfın farklı yazılması (hoodie/HoOdIe), bazı görsellerde etiketin eksik bırakılması veya yanlış sınıflandırma yapılması doğrudan model performansını düşürür. Bu yüzden etiketleme sürecinde sınıf isimleri standartlaştırılmalı, etiketleme kuralları belirlenmeli ve örnek kontroller yapılmalıdır. **Multi-label etiketleme** kullanılıyorsa her etiketin anlamı açık şekilde tanımlanmalıdır.

Segmentasyon veri setlerinde maske kalitesi doğrudan başarıyı etkiler. Maskeler nesne sınırlarına uygun olmalı; taşma, eksik kapsama, kopukluk veya delik gibi hatalar barındırmamalıdır. Maskelerin kalitesi, görsel üzerine **bindirme (overlay) yöntemiyle düzenli olarak kontrol edilmelidir.** Hatalı maskeler düzeltilmeli veya yeniden anotasyon yapılmalıdır.

Son olarak **veri dağılımı** kontrol edilmelidir. Sınıflar arasında aşırı dengesizlik varsa model bazı sınıfları öğrenemez. Bu nedenle her sınıfın yeterli örnek sayısına sahip olması hedeflenmeli, gerekiyorsa veri artırımı **(augmentation)** uygulanmalıdır. Eğitim/validasyon/test ayrımı yapılırken sınıf dağılımı korunmalıdır.

Sonuç olarak **iyi bir veri seti; temiz, çeşitli, dengeli ve doğru etiketlenmiş veriden oluşur** ve başarılı bir model eğitiminin temelini oluşturur.