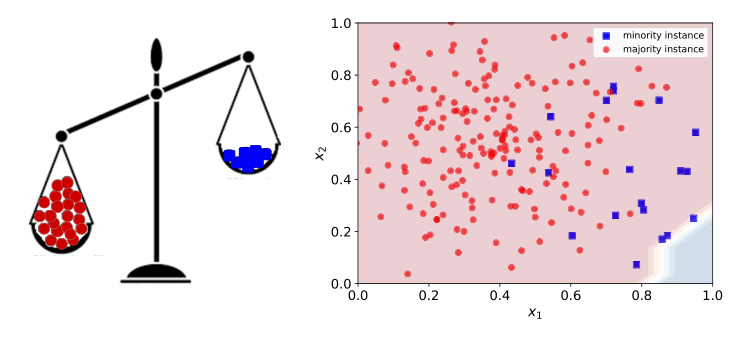
**Dengesiz Veri Setleriyle Makine Öğrenmesi: Gerçek Hayatta Karşılaştığımız Büyük Bir Sessiz Tehdit**

Makine öğrenmesiyle ilgileniyorsan ya da bu alanda bir model geliştirdiysen, "dengesiz veri seti" terimiyle mutlaka karşılaşmışsındır. Göz ardı edildiğinde modelinin başarısını sessizce baltalayabilecek bu durum, özellikle sağlık, finans ve güvenlik gibi alanlarda oldukça kritik sonuçlara yol açabiliyor. Bu yazıda, dengesiz veri setlerinin neden bu kadar önemli bir sorun olduğunu, hangi tekniklerle başa çıkılabileceğini ve gerçek dünyada nasıl etkiler yarattığını konuşacağız.



**Dengesizlik Neden Bu Kadar Büyük Bir Sorun?**

Dengesiz veri setleri, sınıflar arasında ciddi orantısızlık olan veri kümeleridir. Örneğin, bir dolandırıcılık tespit sisteminde dolandırıcılık vakaları %2, geri kalan %98 ise normal işlemler olabilir. Model bu durumda çoğunluk sınıfı (normal işlemler) lehine öğrenir ve azınlık sınıfı (dolandırıcılık) neredeyse yok sayılır.

Başarı gibi görünen %98 doğruluk oranı, aslında büyük bir aldatmacadır. Çünkü model belki de dolandırıcılığı neredeyse hiç tespit edemiyordur. Sağlıkta bu durum çok daha kritik olabilir: Nadir görülen bir hastalığın teşhisi gözden kaçarsa, bu bir insan hayatına mal olabilir.

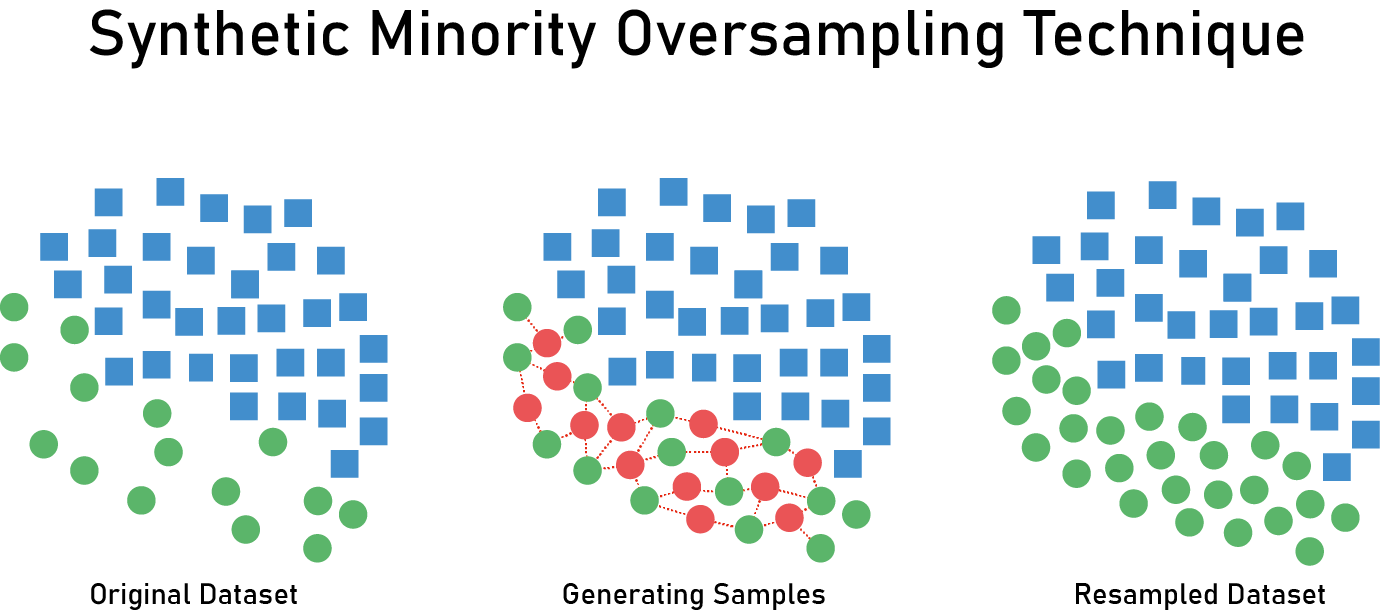
**Bu Sorun Nasıl Aşılır? — Yaklaşımlar ve Teknikler**

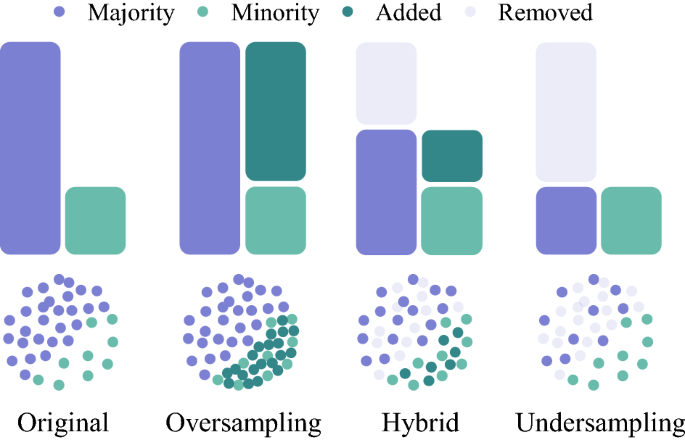
**1. Veri Üzerinden Müdahale: Yeniden Örnekleme**

İlk çözüm yolu veriye dokunmak. Dengesizliği azaltmak için iki temel strateji var:

* **Oversampling (Aşırı örnekleme):** Azınlık sınıfı örneklerini çoğaltarak denge sağlanır. SMOTE (Synthetic Minority Over-sampling Technique), bu alandaki en bilinen yöntemlerden. Var olan azınlık örneklerinden yeni, sentetik veriler türetir.
* **Undersampling (Eksik örnekleme):** Çoğunluk sınıfındaki fazla örnekler azaltılır. Tomek Links gibi yöntemler, bu sınıftaki “gürültülü” verileri temizlemekte oldukça etkili.

Her iki yöntemin de avantajları ve dikkat edilmesi gereken yanları var. Örneğin SMOTE overfitting riskini artırabilir, undersampling ise değerli bilgileri kaybetme riski taşır.





**2. Model Üzerinden Müdahale: Maliyet Duyarlı Öğrenme**

Veriye müdahale etmek yerine modeli daha "duyarlı" hale getirmek de bir seçenek. **Cost-sensitive learning** tam olarak bunu yapıyor: Azınlık sınıfının yanlış sınıflandırılmasının maliyetini artırarak modelin bu sınıfa daha fazla dikkat etmesini sağlıyor.

Bazı algoritmalar buna doğrudan destek veriyor. Örneğin:

* **Gradient Boosting** algoritmalarında sınıf ağırlıklarını ayarlayabiliyorsun.
* **Support Vector Machines (SVM)** gibi modellerde de sınıflar arasında ağırlıklandırma yapılabiliyor.

**3. En Etkili Çözüm: Hibrit ve Ensemble Yaklaşımlar**

Daha güçlü modeller mi arıyorsun? O zaman veri ve algoritma seviyesindeki yöntemleri birlikte kullanmak çok daha etkili sonuçlar verebilir.

* **EasyEnsemble** ve **BalanceCascade**, çoğunluk sınıfını alt örneklemeye tabi tutarak birden fazla sınıflandırıcıdan oluşan modeller kurar.
* **Random Forest**, **XGBoost** gibi ensemble modeller de ROC AUC ve F1 skorları gibi metriklerde yüksek performans gösterir.

Son yıllarda, GAN’lar (Generative Adversarial Networks) ile **sentetik veri üretimi** de yükselişte. Bu yöntemle özellikle az veri bulunan alanlarda (örneğin tıpta nadir hastalıklar) yapay örnekler üretilerek model performansı ciddi şekilde artırılabiliyor.

**Performans Ölçümünde Doğruluk Tuzağına Düşme**

%95 doğruluk seni etkiliyor olabilir, ama bu metrik dengesiz veri setlerinde pek bir anlam ifade etmeyebilir. Daha anlamlı olanlar:

* **Precision (Hassasiyet)**
* **Recall (Duyarlılık)**
* **F1-Score**
* **ROC AUC**

Bunlar, özellikle azınlık sınıfının doğru tahmin edilip edilmediğini çok daha iyi yansıtır. Ayrıca çapraz doğrulamada **stratified sampling** (sınıf oranlarının korunması) kullanmak modelin istikrarını artıracaktır.

**Gerçek Hayattan Örnekler**

* **Finans sektöründe**, dolandırıcılık vakaları %1’den az olabilir. SMOTE ve maliyet duyarlı algoritmalar kullanıldığında bu vakaların yakalanma oranı ciddi şekilde artabiliyor.
* **Tıpta**, nadir görülen bir hastalık için geliştirilen bir teşhis sisteminde, azınlık sınıfı doğru tahmin edilmediğinde çok kritik vakalar atlanabilir. Bu nedenle yeniden örnekleme ve ensemble modeller birlikte kullanılabiliyor.

**Sonuç: Her Problem Aynı Çözümü Gerektirmez**

Dengesiz veri problemi her makine öğrenmesi geliştiricisinin bir noktada karşılaşacağı bir zorluk. Önemli olan, bu durumun farkında olmak ve probleme uygun stratejiler belirleyebilmek.

Kimi zaman yalnızca SMOTE yeterli olurken, kimi zaman maliyet duyarlı öğrenme ve ensemble modellerin bir araya gelmesi gerekebilir. Domain bilgisiyle teknik bilgiyi birleştirebilirsen, modelin sadece doğru değil, aynı zamanda **anlamlı** kararlar veren bir sistem haline gelir.

KAYNAKÇA

<https://ieeexplore.ieee.org/document/5128907>

<https://www.jair.org/index.php/jair/article/view/10302>

<https://link.springer.com/article/10.1007/s10115-022-01772-8>

<https://link.springer.com/book/10.1007/978-1-4419-9326-7>