

Haftalık Rapor

Bu hafta projenin Veri Mühendisliği altyapısının kurulumuna odaklanılmıştır. StatsBomb veri setinin karmaşık ve iç içe geçmiş JSON yapısını çözümleyen, ham veriyi analiz edilebilir formata dönüştüren ETL Pipeline mimarisi tasarlanmıştır. Ayrıca literatür taraması derinleştirilerek modelleme stratejisinde revizyona gidilmiştir.

A. Veri Temini ve Geliştirme Ortamı

- Projenin temel veri kaynağı olan StatsBomb Open Data (Events ve 360-Tracking verileri) lokal çalışma ortamına entegre edilmiştir.
- Python tabanlı veri analizi ortamı hazırlanmış , görselleştirme için mplsoccer ve veri işleme için json kütüphaneleri projeye dahil edilmiştir.

B. Veri İşleme (ETL) Süreci

- Veri setindeki "Olay" (Event) verisi ile anlık oyuncu konumlarını içeren "Freeze Frame" verileri arasındaki ilişki analiz edilmiştir.
- Her bir duran top anında sahadaki 22 oyuncunun konum verisini (x, y koordinatları) ayırtıran ve DataFrame yapısına dönüştüren Parser algoritmasının mantıksal kurgusu tamamlanmıştır.

C. Metodoloji ve Literatür Kararı

- Literatür taraması kapsamında Billings (2024, MIT) tezi detaylıca incelenmiştir. İlgili çalışmada makine öğrenmesi yöntemleri duran top türlerini *sınıflandırmak* amacıyla kullanılmıştır.
- Ancak bu bitirme projesinin özgün değerini artırmak adına, sınıflandırma yerine "Tehdit Değerleme" (Threat Valuation / Regression) yaklaşımının benimsenmesine karar verilmiştir. Amaç olayın türünü bulmak değil, gol olma olasılığını (xT) hesaplamaktır.

D. KARŞILAŞILAN TEKNİK DURUMLAR

- Kaynak verideki oyuncu konumları standart sütunlar yerine, JSON objeleri içinde listeler halinde gelmektedir. Bu durum, doğrudan modellemeyi engellemektedir.
- Bu hafta geliştirilmeye başlanan döngüsel ayrıştırma script'i ile bu verilerin düzleştirilmesi sağlanarak öznitelik mühendisliğine hazır hale getirilmesi hedeflenmektedir.