

Sayıların Ötesine Geçin: Verileri İçgörülere Dönüştürün

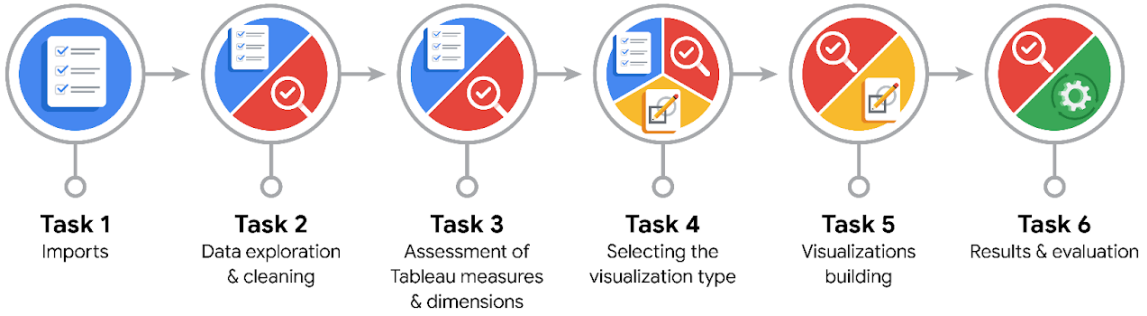


Talimatlar

Bu PACE strateji belgesini, bu ders sonu projesi üzerinde çalışırken kararlarınızı ve düşüncelerinizi kaydetmek için kullanın. Bu belgeyi, veri analiz sürecinin farklı aşamalarındaki yanıtlarınızı ve düşüncelerinizi değerlendirmek için bir kılavuz olarak kullanabilirsiniz. Ayrıca, PACE strateji belgeleri gelecekteki projeler üzerinde çalışırken bir kaynak olarak kullanılabilir.

Referans Kılavuzu

Bu projenin altı görevi vardır; aşağıdaki görsel, PACE aşamalarının bu görevlere nasıl dahil edildiğini belirtir.



Veri Projesi Soruları ve Dikkate Alınması Gerekenler



PACE: Plan Aşaması

- Veri sütunları ve değişkenler nelerdir ve hangileri teslimatınızla en alakalıdır?

Veri Sütunları:

Kimlik & Meta Verileri: video_id, video_duration_sec, verified_status, author_ban_status

İçerik & Moderasyon: claim_status, video_transcription_text

Etkileşim: video_view_count, video_like_count, video_share_count, video_download_count, video_comment_count



Teslimatla En Alakalı Sütunlar:

claim_status → Videonun bir iddia içerip içermediğini belirtiyor.

video_transcription_text → İçeriğin analiz edilmesi için önemli.

video_view_count, video_like_count, video_share_count, video_comment_count → Videoların etkileşim düzeyini gösterir.

author_ban_status → İçerik üreticisinin daha önce yasaklanıp yasaklanmadığını gösterir.

● Değişkenleriniz hangi birimlerde?

- video_duration_sec → Saniye (sn)

- video_view_count, video_like_count, video_share_count, video_download_count, video_comment_count → Sayı (adet)

- claim_status, verified_status, author_ban_status → Kategorik (metin/veri etiketi)

- video_transcription_text → Metin

● Gelecekteki bulgularınızı doğrulamanız veya reddetmeniz gerekeceğini bilerek, EDA'nızı bilgilendirebilecek verilerle ilgili ilk varsayımlarınız nelerdir?

1. İddialı videoların etkileşimi farklı olabilir.
2. Uzun videolar daha fazla izlenebilir.
3. Banlı hesaplar daha fazla iddia alabilir.
4. Doğrulanmış hesaplar daha az iddia alabilir.
5. Yüksek etkileşimli videolar daha çok incelenebilir.

● Eksik veya tamamlanmamış veri var mı?

Evet, claim_status, video_transcription_text, video_view_count, video_like_count, video_share_count, video_download_count ve video_comment_count sütunlarında 298 eksik değer var.



- Bu veri setinin tüm parçaları aynı formatta mı?

Hayır, veri setinde **farklı veri türleri** var:

- **Sayısal veriler** → `int64`, `float64` (ör. video süresi, etkileşim sayıları)
- **Metin verileri** → `object` (ör. `claim_status`, `video_transcription_text`)

Ayrıca, bazı sütunlarda eksik veriler olduğu için format tutarsızlıkları olabilir.

- Bu projeye başlamak için hangi EDA uygulamalarına ihtiyaç duyulacak?

Eksik veri analizi → Eksik değerlerin dağılımını incelemek ve doldurma stratejisi belirlemek.

Özet istatistikler → `describe()` ve `info()` ile veri setinin genel yapısını anlamak.

Veri türü kontrolü → Sütunların doğru formatta olup olmadığını doğrulamak.

Dağılım analizi → Sayısal değişkenlerin histogram ve kutu grafikleriyle dağılımını görmek.

Korelasyon analizi → Değişkenler arasındaki ilişkileri incelemek.

Kategorik değişken analizi → `claim_status`, `verified_status` gibi değişkenlerin frekanslarını incelemek.



PACE: Analiz Aşaması

- Proje hedefine ulaşmak için EDA'nın en etkili şekilde gerçekleştirilebilmesi için hangi adımların atılması gerekiyor?

Eksik verileri işlemek → Doldurmak veya çıkarmak.

Veri türlerini kontrol etmek → Doğru türde olduklarından emin olmak.

Özet istatistikleri incelemek → Ortalama, medyan gibi değerleri görmek.

Veri dağılımlarını görselleştirmek → Histogramlar ve kutu grafikleri kullanmak.

Korelasyon analizi yapmak → Sayısal değişkenler arasındaki ilişkileri keşfetmek.



Kategorik verileri analiz etmek → Frekans dağılımlarını incelemek.

Veri temizliği yapmak → Aykırı ve hatalı verileri düzeltmek.

Veriyi modellemeye hazırlamak → Normalizasyon ve kodlama yapmak.

- EDA birleştirme uygulamasını kullanarak daha fazla veri eklemeniz gerekiyor mu? Bu veri setine filtreleme, sıralama vb. gibi ne tür bir yapılandırma yapılması gerekiyormu?

Veri setine ek veri eklemeye gerek yok. Ancak şunlar yapılabilir:

1. **Filtreleme:** Eksik veya hatalı verileri çıkarmak.
2. **Sıralama:** Videoları video_view_count gibi metriklere göre sıralamak.
3. **Gruplama:** claim_status gibi kategorik verilere göre gruplama yapmak.
4. **Veri Temizleme:** Aykırı değerleri tespit edip temizlemek.

- Hedef kitleye en uygun görselleştirme türleri hakkında ilk varsayımlarınız nelerdir?

Hedef kitle için en uygun görselleştirme türleri şunlar olabilir:

- **Çubuk Grafikler:** claim_status gibi kategorik verileri göstermek için uygun.
- **Dağılım Grafikleri:** video_view_count ile diğer sayısal değişkenler arasındaki ilişkiyi göstermek için kullanışlı.
- **Pasta Grafikleri:** verified_status veya author_ban_status gibi kategorilerin oranlarını görselleştirmek için uygun.
- **Zaman Serisi Grafikleri:** Videoların izlenme sayısı gibi metriklerin zaman içindeki değişimini göstermek için kullanılabilir.



PACE: Yapı Aşaması

- Proje hedeflerini tamamlamak için hangi veri görselleştirmelerinin, makine öğrenimi algoritmalarının veya diğer veri çıktılarının oluşturulması gerekecek?

1. Görselleştirmeler:

- o Çubuk grafikler (kategorik veriler için)
- o Dağılım grafikler (sayısal değişken ilişkisi)
- o Isı haritaları (korelasyonlar için)

2. Makine Öğrenimi:

- o Sınıflandırma modelleri (Logistic Regression, Random Forest)
- o Veri temizleme ve dönüşüm

3. Diğer Çıktılar:

- o Özellik seçimi ve öznitelik mühendisliği.



- Gerekli veri görselleştirmelerinin oluşturulabilmesi için hangi işlemlerin gerçekleştirilmesi gerekiyor?

1. Veri Temizleme: Eksik ve hatalı verilerin düzeltilmesi.
2. Veri Dönüşümü: Sayısal ve kategorik verilerin uygun formatlara dönüştürülmesi.
3. Öznitelik Mühendisliği: Yeni özelliklerin oluşturulması, örneğin, metin verilerini analiz edip anahtar kelimeler çıkarmak.
4. Veri Normalizasyonu: Verilerin aynı ölçeğe getirilmesi (özellikle sayısal veriler için).
5. Filtreleme: İlgisiz veya hatalı verilerin çıkarılması.
6. Gruplama: Kategorik verilere göre gruplama yapılması (örneğin, claim_status).
7. Korelasyon Analizi: Sayısal veriler arasındaki ilişkilerin belirlenmesi

- Bu veri projesindeki görselleştirmeler için hangi değişkenler en uygundur?

Bu veri projesindeki görselleştirmeler için uygun değişkenler şunlar olabilir:

1. claim_status: Kategorik bir değişken, çubuk grafiklerle görselleştirilebilir.
2. video_view_count, video_like_count, video_share_count: Sayısal değişkenler, dağılım grafikleri ve histogramlarla analiz edilebilir.
3. video_duration_sec: Video uzunluğu, sayısal verileri görmek için kullanılabilir.
4. verified_status, author_ban_status: Kategorik veriler, pasta grafikleri ile görselleştirilebilir.
5. video_transcription_text: Metin verileri, kelime bulutları veya metin analizi görselleştirmeleri ile işlenebilir.

- Planlama aşamasına geri dönersek, eksik verilerle (eğer varsa) nasıl başa çıkmayı planlıyorsunuz?

Eksik verilerle başa çıkmak için şu yöntemleri planlıyorum:

1. **Eksik Veriyi İstatistiksel Yöntemlerle Doldurmak:** Ortalama, medyan veya en sık görülen değerlerle doldurma.
2. **Veri Çıkarma:** Eksik verisi fazla olan satırları çıkarma.
3. **İleri Dönük Doldurma:** Eksik verileri komşu değerlerle doldurmak (forward/backward filling).



PACE: Uygulama Aşaması

- EDA ve görselleştirmelerinizden hangi temel içgörüler ortaya çıktı?

EDA ve görselleştirmeler, video etkileşim metrikleri arasındaki ilişkileri, transkripsiyon içeriğinin etkileşimlere etkisini, onay ve yasaklama durumunun rolünü ve eksik verilerin dağılımını gösterdi. Ayrıca, popüler içerik özellikleri ve claim status ile diğer değişkenler arasındaki bağlantılar analiz edildi.

- Oluşturulan görselleştirmeler dayanarak hangi iş ve/veya organizasyonel önerileri öneriyorsunuz?

Görselleştirmelere dayanarak, içerik etkileşimini artırmak için belirli video türlerine odaklanılması, onaylı içeriklerin performansının analiz edilerek strateji belirlenmesi ve eksik verilerin azaltılması için veri toplama süreçlerinin iyileştirilmesi önerilebilir. Ayrıca, claim status ile etkileşim metrikleri arasındaki ilişkiler doğrultusunda içerik yönetim politikaları geliştirilebilir.

- Veriler ve kullandığınız görselleştirmeler hakkında bildikleriniz göz önüne alındığında, başka hangi soruları araştırabilirsiniz?

İçerik etkileşimini artıran faktörler nelerdir? Ban durumu etkileşimleri nasıl etkiliyor? Video süresi ile etkileşim arasında nasıl bir ilişki var? Eksik veriler belirli kategorilerde mi yoğunlaşıyor? Claim status, içerik performansını nasıl etkiliyor?

- Bu görselleştirmeleri farklı kitlelerle nasıl paylaşabilirsiniz?

Görselleştirmeleri farklı kitlelerle paylaşmak için interaktif panolar (Tableau, Power BI), raporlar, sunumlar veya sosyal medya infografikleri kullanılabilir. Teknik kitle için ayrıntılı analiz, iş kitleleri için özetlenmiş içgörüler sunulabilir.