SOSYAL ANKSİYETE VERİ SETİNDE CRİSP-DM SÜRECİ

Sosyal Kaygı Veri Setine Göre İş Hedefi Aşaması:

Problem Tanımı:

Sosyal kaygı (sosyal anksiyete), bireylerin toplumsal ortamlarda yoğun stres, utanma veya kaçınma davranışları yaşamasına neden olan yaygın bir ruhsal bozukluktur. Dünya genelinde milyonlarca insanı etkileyen bu durum, bireyin akademik, sosyal ve profesyonel yaşamını önemli ölçüde olumsuz etkileyebilir.

Bu çalışmanın problemi şudur:

"Davranışsal, yaşam tarzı ve psikolojik değişkenlere bakılarak bireylerin sosyal kaygı düzeyleri (1-10 arasında) doğru şekilde tahmin edilebilir mi?"

İş Hedefi:

Bu problemin çözülmesiyle ulaşılmak istenen iş hedefleri şunlardır:

1. **Erken Müdahale İçin Tahmin Modelleri Geliştirmek:**
   Sosyal kaygı seviyeleri önceden tahmin edilerek bireylere erken psikolojik destek sunulabilir. Bu, hem klinik psikologların karar süreçlerini destekler hem de toplumsal ruh sağlığı yükünü azaltır.
2. **Risk Faktörlerini Ortaya Koymak:**
   Uyku düzeni, stres seviyesi, fiziksel aktivite gibi faktörlerin sosyal kaygıya etkisini analiz ederek risk oluşturan yaşam tarzı seçimleri tespit edilebilir. Bu sayede farkındalık çalışmaları veya kamu politikaları tasarlanabilir.
3. **Kişiselleştirilmiş Müdahale Yaklaşımları Tasarlamak:**
   Farklı birey profillerine göre kaygı seviyelerini tahmin ederek kişiye özel öneriler (örneğin uyku düzeni önerisi, fiziksel aktivite planlaması) yapılabilir.
4. **Eğitim Kurumlarında Kullanım:**
   Gençlerde sosyal kaygı düzeyi yüksek olan öğrencilerin tespiti, rehberlik hizmetlerinin güçlendirilmesini sağlar.

Sosyal Kaygı Veri Setine Göre Veri Anlama Aşaması:

1. Veri Kaynağı:

Veri seti, sosyal kaygı düzeyleri üzerine hazırlanmış sentetik bir veri setidir. Gerçek dünya örneklerine benzer olacak şekilde tasarlanmıştır. Psikolojik anketler, davranışsal ölçümler ve gözlemsel çalışmalar temel alınmıştır.

- Veri seti 10.000+ örnekten oluşmaktadır.
- Eğitim ve araştırma amaçlı kullanılmak üzere oluşturulmuştur.
- Klinik tanı koymak için değil; analiz, modelleme ve içgörü üretimi içindir

2. Veri Setinin Genel Yapısı:

Değişken Kategorileri:

Veri setinde yer alan değişkenler şu başlıklar altında gruplanmıştır:

Kategori	**Örnek Değişkenler**
Demografik	Yaş, Cinsiyet, Meslek
Yaşam Tarzı	Uyku süresi, Fiziksel Aktivite, Diyet Kalitesi, Alkol Kullanımı, Kafein Alımı, Sigara Alışkanlıkları
Sağlık Göstergeleri	Kalp Atış Hızı, Terleme, Baş Dönmesi, Solunum Hızı,
Zihinsel Sağlık	Ailede Anksiyete Öyküsü, İlaç Kullanımı, Terapi Sıklığı
Yaşam Olayları	Yakın zamanda yaşanan büyük olaylar (ör. iş kaybı, taşınma vb.)
Hedef Değişken	Anksiyete Seviyesi (1 ile 10)

3. Veri Tipleri:

- **Sayısal Değişkenler:** yaş, uyku süresi, kalp atış hızı, stres seviyesi, vb.
- **Kategorik Değişkenler:** cinsiyet, sigara kullanımı, ailede anksiyete öyküsü, meslek.
- **Ordinal Değişkenler:** diyet kalitesi, terapi sıklığı gibi sıralı nitelikler.

İstatistiksel Ön Analizler:

Aşağıdaki işlemler gerçekleştirilmiştir

- **describe() komutu** ile sayısal değişkenlerin ortalama, medyan, standart sapma gibi özet istatistikleri çıkarılmış.
- **value_counts()** ile kategorik değişkenlerin dağılımları incelenmiş.
- **Eksik veri analizi** yapılmış (örneğin: isnull().sum() ile eksik sütunlar gösterilmiştir.
- **Korelasyon matrisi ve ısı haritası (heatmap)** çizilmiştir.

- **Boxplot'larla uç değer analizi** yapılmıştır.

Sosyal Kaygı Veri Setine Göre Veri Hazırlama Aşaması

1. Eksik Değer Yönetimi:

Veri setinde eksik değerlerin olup olmadığı analiz edilmiştir.

- df.isnull().sum() gibi yöntemlerle her sütundaki eksik kayıtlar sayılmıştır.
- Eksik verilerin bulunduğu sütunlar için mantıklı stratejiler kullanılmıştır.

Kullanılan Yöntemler:

- **Sayısal değişkenlerde:** Ortanca (median) veya ortalama (mean) ile doldurma işlemi.
    - Örneğin: Uyku süresi, stres seviyesi gibi sütunlar için yapılmıştır.
- **Kategorik değişkenlerde:** En sık görülen değer (mode) ile doldurma işlemi
    - Örneğin: Meslek, ailede anksiyete öyküsü gibi.

Gerekçelendirme: Bu yöntemler, dağılımı bozmadan eksik verileri tamamlamak için yaygın ve geçerli yöntemlerdir.

2. Veri Dönüşümleri:

Veri tipleri ve yapıları modellemeye uygun hale getirilmiştir:

Sayısallaştırma:

- **Kategorik değişkenler** (örneğin: cinsiyet, sigara kullanımı, baş dönmesi, ilaç kullanımı, meslek vb.) LabelEncoder ve OneHotEncoder gibi yöntemlerle sayısal formata çevrilmiş.

Normalizasyon / Ölçekleme:

- Bazı algoritmalar (özellikle KNN, Lojistik Regresyon) için gerekli olan **ölçekleme işlemleri (StandardScaler)** yapılmıştır.
    - Örneğin: Kalp atış hızı, solunum hızı gibi değişkenler normalize edilmiş.

Bu dönüşümler, algoritmaların daha sağlıklı çalışmasını sağlar çünkü farklı ölçeklerdeki veriler model performansını bozabilir.

Değerlendirme:

- Eksik değerler tespit edilip uygun yöntemlerle tamamlanmıştır.
- Kategorik veriler dönüştürülmüştür.
- Ölçekleme yapılmıştır.
- Yeni ve anlamlı değişkenler üretilmiştir.
- Gerekli temizlik ve dönüşümler yapılmıştır.

Sosyal Kaygı Veri Setine Göre Modelleme Süreci:

1. Model Seçimi – Problem Tanımına Uygunluk:

Veri setindeki hedef değişken Anksiyete Seviyesi (1-10) olduğu için:

Seviyeler sınıflandırılıp **(örn. 1–3: Düşük, 4–7: Orta, 8–10: Yüksek)** kategorilere ayrılarak bir **sınıflandırma problemi** yapılmıştır. Kullanılan modeller:

- Logistic Regression
- Random Forest Classifier
- XGBoost Classifier
- KNN
- SVM

2. Model Eğitimi:

Veri eğitim ve test kümelerine doğru bir şekilde ayrılmış:

- train_test_split ile eğitim ve test verisi ayrımı yapılmıştır.
- Genellikle %80 eğitim, %20 test oranı kullanılmıştır.

3. Hiperparametre Ayarlamaları (Model Tuning):

Bazı modeller için hiperparametre ayarlamaları yapılmışır.

- GridSearchCV arama yöntemiyle:
  - Random Forest için n_estimators, max_depth, min_samples_split parametresi kullanılmıştır.

Bu tuning işlemleri, modelin hem doğruluğunu artırır hem de aşırı öğrenme riskini azaltır.

Sosyal Kaygı Veri Setine Göre Değerlendirme Süreci:

1. Doğruluk Ölçütleri

Modelin türüne göre şu metrikler kullanılmıştır:

Sınıflandırma Modelleri için:

- **Accuracy (Doğruluk)**: Genel başarı oranı. Ancak dengesiz veri setlerinde yeterli değildir.

- **Precision / Recall / F1-Score**: Özellikle yüksek anksiyete sınıflarını tahmin etmek için önemlidir.
    - Örneğin: Yüksek kaygı düzeyine sahip bireyleri doğru yakalamak istiyorsak **Recall** ve **F1-score** daha önemlidir.

2. Analiz Derinliği – Yorumlama Kalitesi:

Senin dosyanda model sonuçları sadece sayısal olarak değil, **yorumlarla birlikte** verilmiştir.

- Örneğin, Random Forest modelinin diğer modellere göre daha iyi sonuç verdiği gözlemlenmiş.
- F1-score'ların özellikle dengesiz sınıflarda daha belirleyici olduğu açıklanmıştır.

3. İş Hedefiyle Uyum:

İş hedefin, **bireylerin sosyal anksiyete düzeylerini doğru şekilde tahmin etmek** ve bu bilgiden hareketle erken müdahale olanaklarını geliştirmektir.

- Eğer yüksek riskli bireyler doğru tahmin edilemiyorsa, modelin işlevi zayıflar.
- Bu yüzden yüksek anksiyete sınıflarının doğru tahmin edilmesi hayati önem taşır.

4. Model Karşılaştırması ve Tercih Gerekçesi:

- Birden fazla model denenmiş ve performansları karşılaştırılmıştır.
- Örneğin: Random Forest ve XGBoost modelleri arasında seçim yaparken sadece accuracy değil, **F1-score** ve **genel kararlılık** da dikkate alınmıştır.

Bu yaklaşım değerlendirme sürecinin çok yönlü olduğunu ve yalnızca tek bir metriğe bağlı kalınmadığını gösterilmiştir.

Değerlendirme Sonucu:

Projenin bu bölümü:

- Doğru performans metrikleri seçilmiştir.
- Bu metrikleri derinlemesine analiz edilmiştir.
- İş hedefiyle uyumlu sonuçlara ulaşılmıştır.
- Model karşılaştırmasını mantıklı temellere dayandırılmıştır.

Sosyal Anksiyete Projesine Göre Dağıtım Değerlendirmesi:

1. Streamlit Kullanımı

Evet, projenin son kısmında **Streamlit ile bir kullanıcı arayüzü hazırlanmış** ve modelin çıktısı etkileşimli şekilde sunulmuştur.

Streamlit arayüzünde:

- Kullanıcının yaş, cinsiyet, sigara kullanımı, terapi oturumları gibi verileri elle girebileceği alanlar mevcut.
- "Predict" butonuna basıldığında model çalışıyor ve **anksiyete skoru** veya sınıfı tahmin edilerek kullanıcıya gösterilmiştir.
- Arayüz sade, kullanıcı dostu ve **son kullanıcıyı düşünerek** hazırlanmıştır.

Bu da modelin sadece teknik olarak değil, **pratik olarak da erişilebilir** hale getirildiği gösterilmiştir.

2. Son Kullanıcı Odaklılık

Bu proje, özellikle psikoloji araştırmacıları, danışmanlar veya bireylerin kendi sosyal kaygı düzeylerini anlamaları için kullanılabilir.

Streamlit arayüzü:

- **Teknik bilmeyen** biri için bile basit ve sezgisel bir kullanım sunuyor.
- Tahmin edilen anksiyete düzeyini anlaşılır bir biçimde görselleştirme potansiyeli taşıyor (grafik veya uyarı kutuları gibi).

Bu da, yalnızca veri bilimciler değil, **son kullanıcı** olarak düşünülmüş farklı profillerin de kullanımına uygun bir uygulama olduğunu gösterir.

3. Modelin Arkada Entegre Edilmesi:

Model, Streamlit arayüzüyle başarıyla entegre edilmiştir.

- Eğitimde kullanılan **eğitilmiş model dosyası (.pkl)** yükleniyor.
- Yeni veriler, ön işleme adımlarından geçiriliyor (standardizasyon, eksik değer kontrolü vs.).
- Model bu yeni veriye tahmin yapıyor.

Bu entegrasyon, projeyi sadece analizden ibaret olmaktan çıkarıp **tamamlanmış bir uygulamaya** dönüştürüyor.