**Veri Madenciliği Dönem Proje İsterleri**

**Proje Tanımı**

* Bu proje kapsamında, öğrenciler belirli bir konuyu seçerek **1000+ örnekten ve en az 8 özelliğe sahip** bir veri seti oluşturarak, temizleme ve ön işleme süreçlerini uygulayacak ve **veri madenciliği algoritmalarını kullanarak** analiz yapacaklardır.
* Projede, **veri toplama aşaması ortak yürütülecek**, ancak her grup **bağımsız bir analiz süreci yürüterek** en başarılı modeli geliştirmeye çalışacaktır.
* Her grup **kendi seçtiği en az 5 algoritma ile** model performanslarını karşılaştıracaktır. Sürecin sonunda gruplar, **sonuçlarını detaylı bir rapor ve sunum** ile açıklayacaktır.
* Proje kapsamında yapılacak çalışmalar detaylı şekilde anlatılmalıdır.
* Gruplar belirledikleri konuya göre **ortak veri toplama sürecine** katkıda bulunur.
* Gruplar veri seti belirlendikten sonra kendi çalışma konularını belirler.
* **Veri toplama tamamlandığında tüm gruplar ham veriye erişebilir.**
* Süreç ile ilgili son tarihler duyurular ve ders esnasında bildirilecektir.
* **\*\*\*\*** Her grup en az **10 kişi** olacak şekilde gruplar oluşturulacak, her grubun kendine ait bir proje lideri olacaktır.

**\*\*\*\*** Veri toplama işlemi bitene kadar tüm gruplar ortak çalışacak, bu süreçte proje grupları belirlenerek daha sonra bildirilecek drive dosyasına grup olarak kaydedilecektir.

\*\*\*\***Veri toplama aşamasına katılmayan arkadaşların projeden not alması mümkün değildir. Dolayısı ile 3. Hafta sonunda veri ve proje gruplarının belirlenme işlemi tamamlanmış olmalıdır. Bu tarihten itibaren yapılan başvurular geçersiz sayılacaktır.**

**14 Haftalık Süreç ve Proje Takvimi**

**1. Veri Toplama (Tüm Gruplar Ortak Çalışacak – 1-3. Hafta)**

* Ortak Konu Belirleme: Örneğin "Ekonomik Kriz ve İşsizlik Algısı", "Yapay Zekâ Korkusu", "İklim Değişikliği" gibi gündemdeki bir konu veya çalışılmak istenen herhangi bir konu seçilebilir. Tüm proje gruplarının konu seçimi ve veri toplamada katkıda bulunması gerekmektedir.
* **Veri Kaynaklarının Belirlenmesi:**
  + Twitter, Reddit, haber siteleri, forumlar vb.
  + Grupların belirli anahtar kelimeler ile toplu veri çekmesi.
  + Anket Uygulanması *(Dersi alan öğrenci sayısı 110+, Bölümdeki toplam öğrenci sayısı 1500+)*
  + TÜİK verileri (gelir düzeyi, işsizlik oranı, eğitim seviyesi)
  + Belediye verileri (trafik yoğunluğu, çevre kirliliği, suç oranı)
  + Hava kalitesi ve iklim verileri
  + Sosyal medya ve Google Trends verileri (şehirlerin popülerlik analizi)

**Veri Toplama Stratejisi:**

* + Tüm gruplar aynı yöntemi kullanarak (Python/Web Scraping veya API veya anket hangi yöntem tercih edildiyse) veri çekecektir.
  + Verinin doğruluğunu kontrol etmek için belirli kriterler belirlenecektir.
  + Veri seti **açık kaynaklardan** veya **kendileri tarafından toplanan** verilerden oluşturulmalıdır. **(KAGGLE, UCI GİBİ VERİ TABANLARINDAN ELDE EDİLEN VERİLER KABUL EDİLMEYECEKTİR.)**
  + Veri kaynağı belirtilmeli, **doğrulanabilir olmalı** ve verinin nasıl toplandığı açıklanmalıdır.

**2. Veri Ön İşleme/Dönüştürme**

**(Her Grup Kendi Yolunu Seçecek/Bağımsız Çalışacak – 4-5. Hafta)**

* Eksik verilerin doldurulma veya temizlenme, çıkarılma işlemleri,
* Veri aykırılıkları (outlier detection) belirlenmesi ve düzenlenmesi,
* Veri setinin genel dağılımı ve özet istatistikleri,
* Kategorik değişkenler için encoding işlemleri (One-hot encoding, Label encoding),
* Sayısal veriler için ölçeklendirme (Min-Max Scaling, Standardization),
* Gereksiz veya yüksek korelasyonlu değişkenler belirlenerek çıkarılması,
* Gerekirse veri transformasyonu işlemleri,
* Dağılım grafikleri, histogramlar, korelasyon matrisleri hazırlanması vs. gibi işlemler yapılarak Veri ön işleme aşamaları tamamlanacaktır.

**TÜM BU İŞLEMLER İÇİN HER GRUP, KENDİ VERİ DÖNÜŞÜM TEKNİKLERİNİ SEÇMEKTE ÖZGÜRDÜR. ANCAK AÇIKLANMAYAN İŞLEMLER PUAN KAYBINA SEBEP OLACAKTIR.**

**3. Veri Madenciliği / Makine Öğrenmesi Süreçleri (Her Grup Bağımsız – 6-9. Hafta)**

* Gruplar kendi seçtikleri **en az 5 veri madenciliği algoritmasını** uygular.
* İlk model sonuçları elde edilir ve karşılaştırmalar başlar.
* Model değerlendirme metrikleri hesaplanır (Accuracy, RMSE, Precision, Recall vb.)
* İlk sonuçların neden iyi veya kötü olduğu analiz edilir.
* Gruplar özgünlüklerini gösterebilmeleri için ekstra yöntemler kullanabilirler.
* **İlk yöntemler de, sonrasında yapılan iyleştirmeler de aşama aşama raporda bulunmalıdır.**

**4. Model Optimizasyonu ve İyileştirme (10. Hafta)**

* Hyperparametre ayarlamaları (GridSearch, RandomizedSearch vb.) yapılır.
* Model performansını artırmak için özellik mühendisliği ve farklı dönüşümler denenir.
* Gruplar kendi optimizasyon stratejilerini geliştirerek rekabet avantajı yaratmaya çalışır.
* Sonuçlar kaydedilir ve birlikte tartışmak için grup içi değerlendirme toplantıları yapılır.

**5. Sonuçların Kıyaslanması ve Raporlama (11-12. Hafta)**

* Gruplar en iyi çalışan modelini belirler.
* Seçilen modelin avantajları ve dezavantajları tartışılır.
* Modelin doğruluğu ve genelleme yeteneği değerlendirilir.
* Kıyaslama Kriterleri:
  + En yüksek doğruluk oranı
  + En iyi görselleştirme
  + Modelin açıklanabilirliği
  + Gerçek dünyadaki anlamlılığı (uygulanabilirliği)
* Tüm aşamalar detaylı bir şekilde rapor haline getirilir. Model sonuçları grafikler, tablolar ve istatistiklerle desteklenir. Veri madenciliği sürecinin her adımı belgelenir. Grupların modelleri, optimizasyon süreçleri ve sonuçları kıyaslanır.
* **REKABET ORTAMI OLUŞTURMAK İÇİN GRUPLARIN KENDİ PERFORMANSINI DİĞER GRUPLAR İLE PAYLAŞMAMALARI GEREKMEKTEDİR.**

**6. Sunum ve Değerlendirme (13-14. Hafta)**

* Her grup maximum 15 dakikalık bir sunum yaparak projelerini tanıtmalıdır. Gruplar kendi modellerini hakem (öğretim üyesi + diğer öğrenciler) önünde sunarak ve en iyi model belirlenecektir.
* En iyi modeli bulan veya en anlamlı içgörüler çıkaran ekipler belirlenecektir.
* Projenin teknik doğruluğunun yanı sıra yenilikçilik ve uygulama potansiyeli de değerlendirmeye dahil edilecektir.
* Değerlendirme sonucunda gruplar arası sıralama gerçekleştirilecektir.

**Başarılar Dilerim.**