

## **BİL476/BİL573 – Veri Madenciliği – Dönem Projesi**

**Amaç:** Final projesinde amaç dersimizde öğrendiğiniz tekniklerle ve daha fazlasını kullanarak bir veri madenciliği projesi gerçekleştireceksiniz.

**Notlandırma:** (Projeniz dersteki toplam notunuzun %25'ini oluşturacak)

Proje teklifi: %2 (son gün: 03.06.2025)

Ara rapor: %3 (son gün: 03.07.2025)

Sınıfta sözlü sunum: %5 (son hafta: 21 ve 24 Temmuz)

Proje raporu: %15 (son gün: 31.07.2025)

### **Önemli noktalar:**

Proje konusunda dikkat edilmesi gereken noktalar:

- Proje konusu seçimi çok önemli. Öncelikle ilgili olduğunuz bir alan seçmenizi ve gerçekten işinize yarayabilecek bir proje ortaya koymanızı bekliyorum.
- Projeniz tanımlı ve sınırları belirli bir işi yerine getiriyor olmalı. Yeterli girdi ve çıktı sistematığına sahip olmalıdır
- Grup çalışması olmayacak, herkes kendi projesini yapacaktır
- Proje sunumunuz Powerpoint vb. bir programda hazırlanmalı ve projenizin temel bileşenlerini özetlemelidir. Özellikle amaç, metot, bulgular/sonuçlar, tartışma ve geliştirilebilecek noktalar açıkça belirtilmelidir.
- Sözlü sunuşunuzu sınıfta arkadaşlarınıza yapacaksınız. Açık ve anlaşılır bir dille proje sunumunuzun bileşenlerini anlatmanız bekleniyor. Sözlü sunum 3-4 dakika kadar sürecek. Mümkünse yazdığınız kodla bir demo yapmanız beklenmektedir. Sözlü sunum son hafta ve/veya final haftasında belirlenen bir gün ve saatte yapılacaktır
- Raporunuzun IEEE Konferans Bildirisi formatında ve tercihen 6-10 sayfa arasında olmalı
- Analizleriniz için Python, Java, Matlab, vs. herhangi bir kodlama dilini kullanabilirsiniz. Raporunuza bu kodları da açıklamalarıyla birlikte ekleyiniz. Kodlar ve referanslar sayfa sınırına dahil değildir
- Başka bir ders için verdiğiniz projeyi aynı şekliyle veremezsiniz. Başka bir projenizle benzerlik %30'u geçmemeli. Veri aynıysa yöntemler yeterince farklı olmalı, yöntemler aynıysa veri farklı olmalı
- Uyarı: Literatürde binlerce kez kullanılmış verileri kullanmayın lütfen. Veri çok kullanılmış fakat siz orijinal bir analiz yapacaksanız bu durumda kullanabilirsiniz

### **Proje teklifi:**

- Son gönderme günü 3 Haziran 2025
- Google form üzerinden proje teklifinizi göndereceksiniz. Proje teklifinizde hangi veriyi nasıl kullanacağınızı, temizleyeceğinizi ve analiz edeceğinizi belirtmeniz gerekiyor. Buna veri kaynağı, tanımı ve yapılacak ön analiz ve analizler, olası sonuçlar dahildir.
- Proje teklifinizde ayrıca bir proje takvimi de vermeniz gerekiyor. Takvimde projenizi alt kısımlara ayırıp,

hangi kısmın ne kadar süreceği ve biteceği listelenecek. Bu tamamen uymanız gereken bir takvim değil, duruma göre içerik ve süreleri değiştirebilirsiniz. Bunu daha sonra raporunuzda da belirtmeniz gerekiyor. Takvime uyabildiniz mi, ne kadar uyabildiniz ve neleri neden değiştirmek zorunda kaldınız.

- Sonradan başka bir proje yapmaya karar vererseniz aynı formu tekrar doldurun ve fikir değiştirdiğinizi belirtin

#### **Proje ara raporu:**

- Son gönderme tarihi: 03.07.2025
- Proje ara raporu en fazla 3 sayfa olacak şekilde o zamana kadar ne yaptığının bir özeti olmalı
- Ara raporunuzda son raporda olması gereken 2-3-4. bölümlerin kısaca olması gerekiyor: Proje amaçlarınız, verinin tanımı, kaynağı, değişkenleri ve bu değişkenlerin ölçüm seviyeleri, dağılımlar, kullanılan metotlar, vs. Elinizde sonuçlar varsa bunlardan kısaca bahsedebilirsiniz fakat bunu yapmak zorunda değilsiniz.
- Ara raporda olması gereken bir önemli bir bilgi de proje takvimi ve o zamana kadarki bu takvime uyum durumunuz. Ayrıca projenizi bitirmek için neler kaldı ve ne gibi bulgular/sonuçlar bekliyorsunuz sorusuna da yanıt vermeniz gerekiyor.

#### **Proje sunumu:**

- Sözlü sunumlar dönemin son haftasında veya duruma göre final haftasında uygun bir günde yapılacak
- Sözlü sunum için bir Powerpoint vs. sunum dosyası hazırlamanız ve sınıfta arkadaşlarınıza canlı olarak sunmanız gerekiyor
- Sunumunuz 3-4 dakika uzunlukta kısa bir yıldırım sunum olacak
- Sunum dosyanız raporunuzda olması gereken aşağıdaki plana uygun bir şekilde hazırlanmalı
- Size gönderilecek şablonu kullanmanız gerekiyor
- Slayt sayısı konusunda bir sınırlama yok fakat 3-4 dakikaya sığabilmesi için 5-6 slayttan fazla olmamalı

#### **Proje raporu:**

Proje raporu ve sunumuzdan olması gereken kısımlar aşağıda verilmiştir. Ayrıca size bunlar için şablon da gönderilecek. Proje raporu son gönderme tarihi 31.07.2025.

#### **1. Özet (Abstract)**

#### **2. Giriş (Introduction)**

- Motivasyon (neden bunu yapıyorsunuz?)
- Literatür taraması: Bu veri daha önce kullanıldı mı? Kullanıldıysa neler yapıldı ve ne sonuçlara ulaşıldı? Siz bunlardan farklı olarak ne yaptınız?
- Genel metodoloji (Veri ön işleme, işleme ve analizi için hangi metotları hangi gerekçelerle kullandınız? Bu metotları çok kısaca tanıtınız. Orijinal çalışmaları referans olarak gösteriniz)
- Amaçlarınız (çok açık bir şekilde madde madde yazılmalı)

#### **3. Veri seti, veri özellikleri, öznitelikler (Data)**

- Veri kaynağı
  - o URL linki, siz mi topladınız (“in-house data”), vb. Ne zaman, nerede, hangi yöntemlerle toplanmış? Ne tür bir veri: yapılandırılmış, yapılandırılmamış, hibrit, vs.
- Ön işleme aşamaları: Burada eğer hazır veri seti kullanmıyorsanız (veya kullandığınız hazır veri seti ham veri seti ise) yaptığınız veri işleme işlemlerini açıklayın.
  - o (varsa) görüntü işleme aşamaları
  - o Ses işleme
- Öznitelik (değişkenlerin) analizi
  - o Hatalı veriler, eksik veriler var mı? Varsa bunlar için ne yaptınız?
  - o Her bir değişkeninizin ölçüm seviyeleri verilmeli (sınıflayıcı, sıralayıcı, aralıklı ve oransal şeklinde. Ya da metin, resim, ses dosyası vs. olarak isimlendirin)
  - o Öznitelikler arasındaki ilişkiler:
    - Öznitelikler arasında korelasyonlar
    - Özniteliklerin sonuç ile korelasyonu
    - Veri dağılımlarının istatistiksel parametreleri (ne tip bir dağılım: Normal, durağan, Poisson, vb.)
    - Normal dağılım durumunda Kovaryans matrisi
    - Sayısal olmayan fakat gruplanabilen değişkenler için frekanslı çubuk grafikleri
- Kelimeler, tweetler, metinler, vb. kullandıysanız
  - o Öznitelik seçimi veya düzenlemesi varsa bu işlemlerin detayları (PCA Principal Component Analysis) veya ICA (Independent Component Analysis)
- Sınıf veri dağılımları bilgileri: Data balanced / imbalanced ? Sınıflardaki veri sayılarını eğitim verileri için dengele
- Data normalizasyonu, öznitelik normalizasyonu (lineer, normal, log, özel bir dönüşüm, vb.) yapıldıysa, neden ve nasıl yaptınız?
- Verinin görsel olarak 2D gösterimi (sınıf değerleriyle beraber), multidimensional data visualization
- Veri görselleştirmesi, işlem sonrası verilerin veya örüntülerin gösterimi, vb.

#### 4. Kullanılan metodoloji (Methods)

- Kullanılan metodoloji, modeller ilgili kısa tanımlar, yaklaşımlar, literatür bilgisi
- Neden bu metod/modeller seçildi?

#### 6. Bulgular (Results)

- Benzerlik, farklılık, doğruluk, vb. Performans metrikleri
- Modeller arası karşılaştırma tabloları ve grafikleri
- Değişik modeller ve veriler arasında istatistiksel doğruluk ve önemlilik/anlamlılık testleri (t testi, ANOVA, Ki kare testi, vs.)

#### 7. Sonuçlar ve Tartışma (Conclusions and Discussion)

- Bu projede kısaca ne yaptığımızın özeti. Hangi model neden daha iyi? Açıklama?
- Proje sonucunda ne öğrendik? Kazancımız nedir? Katkımız ne oldu?
- Projede yap(a)madıklarımız neler var? Bunların sebepleri neler? (Kısıtlamalar/Limitations)
- Gelecek çalışmalar neler olabilir? (Future Work)

#### 8. Referanslar

Not: Rapor uzunluğu IEEE konferans bildiri formatında (Word ve Latex) 6-10 sayfa arasında olacaktır. Buna ek olarak Ekler (Appendices) adı altında bir bölüm daha oluşturup bunu bildiriden ayrı olarak verebilirsiniz. Appendix içeriğinde özellikle Önişleme, Öznitelik seçimi, korelasyonlar, vb. gibi bilgilerin yanı sıra program kodu da konabilir.