2025-2026 Güz Yarıyılı Makine Öğrenmesi Dersi 2. Ödevi

Konu: Öğrenme Modellerinde Özellik Seçimi

Problem: Bu ödevde, bir denetimli öğrenme (supervised learning) probleminde özellik seçiminin sistem performansına etkisi değerlendirilecektir. Bunun için *Online News Popularity (UCI Machine Learning Repository)* veri kümesi kullanılarak bir haberin **popüler olup olmayacağını tahmin eden** bir ikili sınıflandırma modeli geliştirilecektir.

Online News Popularity (UCI Machine Learning Repository) veri kümesi Mashable tarafından yayınlanan 39.644 haber makalesinin istatistik bilgilerini içermektedir. Veri kümesine aşağıdaki linkten erişebilirsiniz: **https://www.kaggle.com/datasets/thehapyone/uci-online-news-popularity-data-set**

İşlem Adımları:

1. **Veri Ön İşeme:**
   a. Ayırt edici bilgi içermeyen **url** gibi özellikleri çıkarın.
   b. Hedef değişkeni (shares) için medyan değerine (1.400) göre 1400'ü eşik seviyesi olarak alıp veri kümesindeki örnekleri **shares >= 1400 ise 1, değilse 0** olarak etiketleyin.
   c. Veriyi %80 eğitim, %20 test olacak şekilde bölün.
   d. Eğitim verisi üzerinde *5-fold cross validation* yöntemini kullanın. Bu durumda her adımda 4 parçayı eğitim, 1 parçayı validasyon için kullanın.

2. **Özellik Seçimi Yöntemlerinin Uygulanması**
   Aşağıdaki üç farklı yöntemi kullanarak en önemli özellikleri belirleyin:
   a. **Filtreleme Yöntemi (Filter Method):**
   Pearson Korelasyonu kullanarak birbiriyle yüksek korelasyona sahip özellikleri eleyin ve en iyi 15 özelliği seçin.
   b. **Sarmalayıcı Yöntem (Wrapper Method):**
   Recursive Feature Elimination(RFE) yöntemini Lojistik Regresyon algoritması ile birlikte kullanarak en iyi 15 özelliği belirleyin.
   https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html
   c. **Gömülü Yöntem (Embedded Method):**
   Gömülü yöntem olarak Rasgele Orman(Random Forest) yöntemini kullanın. Özellik önem skoruna göre ilk 15 özelliği seçin.
   https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

3. **Performans Ölçümü:**
   a. Veri kümesindeki eğitim örneklerini önce bütün özellikler için, daha sonra (a,b,c) maddelerinde elde edilen özellikler ile **Lojistik Regresyon** yöntemi ile k-fold cross validation kullanarak ayrı ayrı eğitiniz. Aşırı öğrenme varsa iyileştirme yapınız. Eğer iyileştirme ihtiyacı varsa nasıl yaptığınızı raporunuzda anlatınız. Daha sonra %20 örnek ile test ederek test başarısını hesaplayınız.
   b. (a) şıkkında en başarılı sonucu veren yöntem hangisiyse (örneğin RFE) o yöntemi kullanarak bu sefer en iyi 15 özelliği değil, **sistem başarısını en yüksek veren özellikleri seçiniz**. Bu özellikleri kullanarak modeli yine **Lojistik Regresyon** yöntemi ile (a) şıkkındaki gibi eğitiniz. Daha sonra %20 örnek ile test ederek test başarısını hesaplayınız.

c. (a ve b) şıklarında belirtilen işlemleri yaparak test sonuçlarını aşağıda verilen tabloya yerleştiriniz.

Yöntem	Özellik Sayısı	Doğruluk(Accuracy)	F1-Skoru	Eğitim süresi
Tüm Özellikler	59			
Filtreleme	15			
Sarmalayıcı	15			
Gömülü	15			
?? Yöntemi	??			

d. En başarılı yöntem için karışıklık matrisini(confusion matrix) veriniz.
e. Bütün yöntemler için seçilen özellikleri bir tabloda veriniz. Hangi özelliklerin ortak olduğunu da gösteriniz.

4. Sonuç:
Elde ettiğiniz sonuçları yorumlayan bir paragraf yazınız.

Değerlendirme:

Ödevler %60 kod ve %40 rapor olmak üzere 100 puan üzerinden değerlendirilecektir.

Rapor (40 Puan)

- Rapor, IEEE makale formatında hazırlanmalıdır.
- Rapor dili bilimsel makale dili olmalıdır. Raporda görseller ve tablolar kullanılmalı ve yapılan değerlendirmeler olabildiğince kaynaklarla desteklenmelidir.
- Metin içerisinde görsellere ve tablolara kesinlikle referans verilmelidir.
- Yazım hataları, zaman kullanımı ve dil bilgisi kontrol edilmelidir.
- Rapor içeriği aşağıdaki ana bölümleri içermelidir:
  - Başlık: Ödev konusu, Ad, soyad ve numara (Makale başlığı formatında yazılmalı)
  - Özet: Yaptığınız çalışmayı, elde ettiğiniz sonuçları bir paragraflık özet olarak veriniz.
  - Giriş: Ödev konusunu tanıtan bir paragraflık bir giriş yapınız. Problemin tanımı, çalışmanın amacı, çalışmanın kapsamı hakkında bilgi veriniz.
  - Deneysel Analiz: Yukarıda işlemler bölümünde belirtilen işlemler için elde ettiğiniz sonuçları tablo ve istenilen grafikleri kullanarak gösteriniz. Başarı değerlerini eğitim, doğrulama ve test için ayrı ayrı veriniz.
  - Sonuç: Elde ettiğiniz sonuçları bir paragrafta yorumlayınız.

Kod (40+20 Puan)

Ödevdeki tüm işlemler için hazır kütüphanelerden faydalanabilirsiniz.

Kullanılan Kütüphaneler ve Gerekli Dosyalar

Kodlar TensorFlow, Keras, Scikit-learn kullanılarak Python ortamında geliştirilmelidir.
ReadMe.txt dosyası içerisinde kodların nasıl çalıştırılabileceği kısaca açıklanmalıdır.

Örnek ReadMe.txt:

```
## Açıklama
Makine Öğrenmesi (BLM5110) dersi kapsamında TensorFlow ve Keras kullanılarak geliştirilen sınıflandırma modeli.

## Gereksinimler
pip install -r requirements.txt

## Çalıştırma
Eğitim:
python train.py
Değerlendirme:
python eval.py

### Dosya Düzeni
proje_klasoru/
  |- train.py
  |- eval.py
  |- requirements.txt
  |- dataset/
  |- results/
```

Gerekli kütüphaneler ve versiyonları requirements.txt dosyasında belirtilmelidir (Anaconda command prompt ile environment içerisindeyken 'conda list' komutu kullanılarak elde edilebilir).
ReadMe.txt ve requirements.txt dosyası eksikliğinde her birinden 10 puan kırılacaktır.

Kodun Modülerliği (20 Puan)

- Kodlar sadece tek bir main fonksiyonundan oluşmamalı, modüler bir yapı izlenmelidir.
- Gerektiğinde kod, çeşitli scriptlere bölünmelidir (train.py, dataset.py, eval.py gibi).
- Kodun okunabilirliği ve anlaşılabilirliği açısından fonksiyon ve sınıflar kullanılmalıdır.
- Fonksiyon ve sınıflar, anlamlı isimlendirmeler ile tanımlanmalı ve kodun anlaşılabilirliği sağlanmalıdır.
- Fonksiyonlar docstring veya yorum satırları kullanılarak işlevleri ve parametreleri açıklanmalı.

Bir fonksiyon için örnek bir docstring:

```
def complex(real=0.0, imag=0.0):
  """Bir karmaşık sayı oluştur.
  Argümanlar:
  real -- gerçek kısım (default 0.0)
  imag -- sanal kısım (default 0.0)
  """
  if imag == 0.0 and real == 0.0:
    return complex_zero
```

Kodun Çalışabilirliği ve Testler (40 Puan)

- Kodun hatasız şekilde ilgili ödevde istenen çıktıları sağladığı ve çalışabilir olduğu test edilmelidir.
- Ara sonuçlar (epoch eğitim/validasyon loss bilgileri vb.) ve sonuçlar konsol üzerinde veya .txt formatında kaydedilerek sunulmalıdır.
- Çalışmadan alınan sonuçların yeterli bir başarı elde etmesi gerekmektedir.

Önemli Not1: Yazdığınız kod ve raporunuz bir başka öğrencininkine veya internetteki kaynaklara belli bir orandan fazla benzerse veya chatGPT gereğinden fazla kullanılırsa kopya olarak değerlendirilecektir.

Önemli Not2: Sisteme ödev dosyalarınızı yükleyiniz. Link bırakmayınız. Link olarak bırakılan ödevler değerlendirilmeyecektir.

Ödev Teslimi: Ödevinizi **5 Ocak 2026 Pazartesi 23.45'e** kadar sisteme yükleyiniz. Lütfen sistemde olabilecek aksaklıkları da hesaba katarak ödevinizi sisteme yükleme işlemini son dakikalara bırakmayınız. E-mail ile gönderilen ödevler kabul edilmeyecektir.