

[**Mühendislik ve Mimarlık Fakültesi**](https://www.konya.edu.tr/muhendislikvemimarlik)

[**Bilgisayar Mühendisliği Bölümü**](https://www.konya.edu.tr/bilgisayarmuhendisligi)

**Bilgisayar Mühendisliği Uygulama Tasarımı Dersi Proje Formu**

|  |
| --- |
| **Proje Başlığı** |
| Asetilkolinesteraz (AChE) Enzimi İnhibitörlerinin Makine Öğrenmesi ile Tahmin Edilmesi |

|  |  |
| --- | --- |
| **Öğrenci Bilgileri** | |
| **Öğrenci No** | 22100011053 |
| **Öğrenci Ad Soyad** | Nisa AKAR |

|  |
| --- |
| **Dr. Öğr. Üyesi**  **Özlem ERDAŞ ÇİÇEK** |

**Şubat 2025**

**Konya**

**İçindekiler**

[1. Proje Konusunun Önemi ve Özgün Değeri 1](#_Toc51147498)

[2. Projenin Amacı ve Hedefi 1](#_Toc51147499)

[3. Projenin iş-zaman çizelgesi 2](#_Toc51147500)

[4. Projede kullanılan donanımlar ve yazılımlar ile ilgili bilgileri 4](#_Toc51147501)

[4.1. Donanımlar 4](#_Toc51147502)

[4.2. Yazılımlar 4](#_Toc51147503)

[5. Projenin yapım aşamaları 4](#_Toc51147504)

[Kaynaklar 4](#_Toc51147505)

# Proje Konusunun Önemi ve Özgün Değeri

Asetilkolinesteraz (AChE) enzimi, nörolojik hastalıkların tedavisinde önemli bir hedef olup, Alzheimer gibi nörodejeneratif hastalıkların tedavisinde kullanılabilecek potansiyel ilaçların tasarımında büyük bir öneme sahiptir. Alzheimer hastalığı, dünya çapında yaygın bir hastalık olup, etkili tedavi yöntemlerinin bulunması büyük bir araştırma alanı oluşturmuştur. AChE inhibitörleri, bu enzimi engelleyerek sinir hücreleri arasında iletişimin düzgün bir şekilde gerçekleşmesini sağlar ve böylece Alzheimer hastalığının semptomlarını hafifletebilir. Ancak, AChE inhibitörlerinin etkinliğini tahmin etmek ve doğru ilaçları geliştirmek oldukça karmaşık bir süreçtir. Bu nedenle, yeni ilaç keşif süreçlerinin hızlandırılması ve etkinliğinin artırılması için modern teknolojiler, özellikle makine öğrenmesi, büyük bir potansiyele sahiptir.

**Özgün Değer**

Günümüzde, AChE enzimi üzerinde yapılan çalışmaların çoğu sınırlı sayıda molekülle yapılmakta ve mevcut tedavi yöntemlerini geliştirebilmek için daha fazla bilgiye ihtiyaç duyulmaktadır. Proje, geniş veri setleri kullanarak, Alzheimer tedavisinde kullanılabilecek AChE inhibitörlerini daha doğru bir şekilde tahmin etmeyi amaçlamaktadır. Makine öğrenmesi yöntemlerinin, özellikle derin öğrenme, rastgele ormanlar ve destek vektör makineleri gibi algoritmaların, bu projede kullanılması, AChE inhibitörlerinin etkili bir şekilde tahmin edilmesini sağlayarak, yeni ilaçların geliştirilmesine katkı sağlayacaktır. Özgünlük, bu alanda kullanılacak olan modellerin doğruluğunun artırılması, verilerin kapsamının genişletilmesi ve yenilikçi bir yaklaşım sunulmasıyla sağlanacaktır.

# Projenin Amacı ve Hedefi

Araştırma kapsamında, Asetilkolinesteraz (AChE) inhibitörlerinin etkinliğini tahmin etmek amacıyla moleküler tanımlayıcılar, parmak izi yöntemleri ve makine öğrenmesi algoritmaları kullanılarak bir model geliştirilecektir. Geliştirilen modelin, literatürde mevcut olan yöntemlere göre daha yüksek doğruluk oranlarına sahip olması, etkili ve hızlı bir şekilde çalışabilmesi, aynı zamanda AChE inhibitörlerinin tahmininde doğru sonuçlar elde edilmesi hedeflenmektedir.

Projenin bir diğer önemli hedefi, modelin çevrimiçi bir şekilde çalışabilmesi ve çeşitli biyolojik verilerle uyumlu hale getirilerek potansiyel ilaç adaylarını hızlı bir şekilde tarayabilmesidir. Bu süreç, Alzheimer gibi nörolojik hastalıkların tedavisinde kullanılabilecek yeni moleküllerin keşfine katkı sağlayacaktır.

Araştırma tamamlandığında, elde edilen modelin biyoteknoloji ve ilaç keşfi alanında kullanılması amacıyla ilgili akademik ve endüstriyel iş birlikleri için başvurular yapılacaktır. Bu sayede, geliştirilen modelin uluslararası alanda da etkin bir araç olarak kullanılabilmesi sağlanarak, ülkemizdeki ilaç araştırmaları ve geliştirme süreçlerine önemli katkılar sağlanması amaçlanmaktadır.

# Projenin iş-zaman çizelgesi

Projede yer alacak başlıca iş paketleri (İP) ve hedefleri, her bir iş paketinin hangi sürede gerçekleştirileceği, başarı ölçütü ve projenin başarısına katkısı “İş-Zaman Çizelgesi” doldurularak verilir.

İş-zaman çizelgesi

|  |  |  |  |
| --- | --- | --- | --- |
| İP.  No. | İş Paketlerinin Adı ve Hedefleri | Zaman Aralığı (Haftalık) | Başarı Ölçütü |
| 1 | **Moleküler Veri Setlerinin Hazırlanması:** AChE inhibitörlerini tahmin etmek için kullanılacak geniş moleküler veri setlerinin toplanması ve tüm veri setlerinin ortak bir formata dönüştürülmesi. Moleküler veriler arasında tutarlılığı sağlamak ve tek tip bir giriş verisi oluşturmak. | 1. hafta | Veri setlerinin analiz edilerek, ortak formatta toplanması ve dönüştürülmesi. Başarı ölçütü: Ortak formatta minimum %95 doğrulukla veri setlerinin dönüştürülmesi. Araştırmaya katkı: %10 |
| 2 | **Kimyasal ve Biyolojik Özelliklerin Çıkarılması:** Moleküler tanımlayıcılar ve parmak izi yöntemleriyle, AChE inhibitörlerinin biyolojik ve kimyasal özelliklerinin çıkarılması. Kimyasal yapılar arasındaki ilişkiyi daha iyi anlamak için farklı biyolojik ve kimyasal özelliklerin analiz edilmesi. | 2. ve 3. haftalar | Moleküllerin biyolojik ve kimyasal özelliklerinin doğruluğu ve anlamlılığı. Başarı ölçütü: Özelliklerin %90'ının doğru çıkarılması ve literatürle uyumlu olması. Araştırmaya katkı: %15 |
| 3 | **Veri Normalizasyonu:** Moleküler veri kümelerindeki kesikli ve sürekli değerlerin normalizasyonu. Kimyasal verilerde farklı değer aralıklarının ve ölçüm birimlerinin olması nedeniyle veri normalizasyonunun yapılması. | 4., 5. ve 6. haftalar | Normalizasyonun doğru yapılması, veri setinin eğitim için uygun hale getirilmesi. Başarı ölçütü: Veri setinin %95'inin doğru şekilde normalize edilmesi. Araştırmaya katkı: %10 |
| 4 | **Özellik Seçimi ve Dimensionalite Azaltma:** Kimyasal ve biyolojik özelliklerden AChE inhibitörleri ile ilişkili olanlarının seçilmesi. Gereksiz verilerin ve özelliklerin çıkarılması, modelin verimli çalışması için önemli olan özelliklerin belirlenmesi. | 7., 8. ve 9. haftalar | Modelin doğruluğunu artırmak için uygun özelliklerin seçilmesi. Başarı ölçütü: Özellik seçimi ile model doğruluğunun %10 artması. Araştırmaya katkı: %15 |
| 5 | **Makine Öğrenmesi Modellerinin Seçilmesi ve Eğitilmesi:** Derin öğrenme, rastgele ormanlar ve destek vektör makineleri gibi makine öğrenmesi algoritmalarının seçilmesi ve AChE inhibitörlerinin tahmin edilmesi için eğitim verilmesi. Model parametrelerinin optimize edilmesi, çapraz doğrulama ile algoritmaların performansının artırılması. | 10., 11. ve 12. haftalar | En uygun algoritmanın seçilmesi ve eğitim başarısının artırılması. Başarı ölçütü: Modelin doğruluğunun %20 artması ve overfitting’in önlenmesi. Araştırmaya katkı: %30 |
| 6 | **Model Doğrulama ve Test:** Eğitim verisi dışında tutulmuş verilerle modelin test edilmesi, modelin doğruluğunun ölçülmesi. Ayrıca, modelin gerçek dünyadaki verilerle test edilmesi için küçük çaplı denemeler yapılması. | 13. ve 14. haftalar | Test aşamasında model doğruluğunun artması, gerçek dünyada kullanılabilirliğinin test edilmesi. Başarı ölçütü: Test verisi doğruluğunun %90’ın üzerine çıkması. Araştırmaya katkı: %25 |

# Projede kullanılacak yöntem, donanımlar ve yazılımlar ile ilgili bilgiler

Projenin yazılım ve donanım kısmında Python programlama dili kullanılacaktır. Verilerin işlenmesi, özelliklerin çıkarılması ve makine öğrenmesi modellerinin geliştirilmesi için Numpy, pandas ve scikit-learn gibi kütüphaneler tercih edilecektir. Bu kütüphaneler, verilerin hazırlanmasında ve analiz edilmesinde önemli rol oynayacak ve makine öğrenmesi algoritmalarının uygulanmasını sağlayacaktır. Ayrıca, RDKit gibi kimyasal veri işleme araçları kullanılarak moleküler parmak izlerinin çıkarılması işlemi yapılacaktır. Bu yazılım araçları, proje hedeflerinin gerçekleştirilmesinde gerekli olan verimliliği ve doğruluğu sağlayacaktır.

# Projenin yapım aşamaları:

1. **Hafta:** **Proje Konusu Belirlenmesi ve Planlama**

Bu hafta, proje kapsamında ele alınacak temel problem tanımlanmış ve çözüm yöntemi olarak makine öğrenmesi temelli bir yaklaşım benimsenmiştir. Özellikle Alzheimer hastalığında kritik rol oynayan Asetilkolinesteraz (AChE) enzimi hedef alınarak, bu enzimi inhibe eden bileşiklerin biyolojik etkinliğini (bioactivity) tahmin etmek amaçlanmıştır. Bu kapsamda, projede kullanılacak veri kaynakları, modelleme stratejileri ve değerlendirme metrikleri üzerine kapsamlı bir literatür taraması gerçekleştirilmiştir.

Gerçekleştirilenler:

* Projenin amacı net bir şekilde belirlendi: Alzheimer hastalığına yönelik AChE inhibitörlerinin tahmini.
* PubChem, ChEMBL gibi kaynaklardan ulaşılabilir biyolojik aktivite verileri incelendi.
* AChE enziminin biyokimyasal işleyişi ve inhibitörlerle etkileşimi hakkında akademik kaynaklardan bilgiler edinildi.
* Kullanılabilecek makine öğrenmesi algoritmaları (örneğin, Random Forest, XGBoost, SVM) analiz edildi.
* Veri toplama, ön işleme, model eğitimi, doğrulama ve test aşamalarından oluşan genel proje yol haritası oluşturuldu.
* Araştırma sırasında kullanılan makaleler kaynakça notlarına eklendi.

1. **Hafta:** **Veri Setinin Hazırlanması ve Eksik Verilerin İncelenmesi**

Bu hafta, AChE inhibitörlerine ait moleküler yapıların ve bu yapıların biyolojik aktivite değerlerinin yer aldığı veri seti elde edildi. İlk adım olarak, veri setinin yapısı analiz edildi ve eksik değerlerin varlığı kontrol edildi. Verinin analiz edilebilir hale gelmesi için eksik veya hatalı gözlemler belirlendi ve uygun veri temizleme teknikleri uygulandı. Ayrıca, verideki kategorik ve sayısal değişkenler ayrı ayrı ele alınarak uygun doldurma stratejileri benimsendi.

Gerçekleştirilenler:

* Veri seti .csv formatında projeye dahil edildi ve Python ortamında (pandas) analiz edildi.
* Eksik veriler tespit edildi; özellikle biyolojik aktivite sınıflarındaki eksikliklerin model performansını olumsuz etkileyebileceği göz önünde bulundurularak önlem alındı.
* Sayısal kolonlarda (örneğin: moleküler ağırlık, logP) eksik değerler medyan veya ortalama ile dolduruldu.
* Kategorik kolonlarda (örneğin: inhibitör tipi) eksik veriler mod değeri kullanılarak tamamlandı.
* Bu adımlar sayesinde veri kümesindeki istatistiksel dağılım bozulmadan korunmuş oldu.
* Veriler normalize edilmeye ve özellik mühendisliği adımlarına hazırlanacak şekilde hazırlandı.

metin, ekran görüntüsü, yazı tipi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Şekil 1.**Eksik Veri Sayısı Grafiği

1. **Hafta:** **Kimyasal ve Biyolojik Özelliklerin Çıkarılması**

Bu hafta, moleküler yapılar üzerinde hesaplamalı kimya yöntemleriyle çeşitli kimyasal ve biyolojik özellikler (deskriptorlar) çıkarılmıştır. Bu özellikler, makine öğrenmesi algoritmalarının eğitimi sırasında modelin öğrenebileceği anlamlı girdiler sağlamak amacıyla elde edilmiştir. RDKit kütüphanesi kullanılarak her molekül için yapısal ve fizikokimyasal özellikler hesaplanmış ve ayrıca moleküllerin benzerliklerini belirlemek amacıyla parmak izi (fingerprint) vektörleri oluşturulmuştur.

Gerçekleştirilenler:

* RDKit kütüphanesi ile aşağıdaki özellikler hesaplandı:
  + MolWt (Moleküler Ağırlık): Molekülün toplam atom kütlelerinin toplamı.
  + LogP (Hidrofilik/Hidrofobik Dağılım Katsayısı): Molekülün yağda çözünürlüğünü belirler.
  + HBA (Hidrojen Bağlayıcı Alıcı) ve HBD (Hidrojen Bağlayıcı Donör) sayıları: Molekülün hidrojen bağı yapma kapasitesini ifade eder.
  + TPSA (Toplam Polar Yüzey Alanı): Molekülün polarlık düzeyini yansıtır ve biyoyararlanımda önemlidir.
* Moleküllerin yapısal benzerliklerini sayısal olarak ifade edebilmek amacıyla Morgan Fingerprint yöntemiyle 1024-bit uzunluğunda ikili vektörler üretildi. Bu vektörler, molekülün atomik çevre yapılarının ikili temsilleridir.
* Tüm çıkarılan özellikler tek bir veri çerçevesinde birleştirildi ve makine öğrenmesi modellerine giriş verisi olarak kullanılmak üzere tablo formatına getirildi.

metin, ekran görüntüsü, yazı tipi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Şekil 2**. Moleküllerin Kimyasal ve Biyolojik Özellikleri

1. **Hafta**: **Veri Setinin Normalizasyon İçin Hazırlanması**

Bu haftada, moleküler özelliklerin farklı sayısal ölçeklerde bulunması nedeniyle veri setinde normalizasyon ihtiyacı ortaya çıkmıştır. Özellikle MolWt özelliğinin 100–900 arasında, LogP'nin -3 ila 10 arasında ve TPSA'nın 0–200 arasında değiştiği gözlemlenmiştir. Bu tür ölçek farklılıkları, modelin bazı özellikleri fazla önemseyip bazılarını ihmal etmesine neden olabileceğinden, tüm özelliklerin aynı ölçekte temsil edilmesi gerekliliği doğmuştur. Bu amaçla çeşitli ölçeklendirme yöntemleri karşılaştırılmıştır.

Yapılan değerlendirmeler sonucunda:

* Min-Max Scaling ve Z-Score yöntemleri analiz edilmiştir.
* Uç değerlere karşı daha dayanıklı olması nedeniyle Z-Score standardizasyonu tercih edilmiştir.
* Bu yöntemle her özelliğin ortalaması sıfır, standart sapması bir olacak şekilde yeniden ölçeklendirilmesi hedeflenmiştir.

Böylece, modelin daha dengeli ve sağlıklı öğrenme yapması için önemli bir zemin hazırlanmıştır.

1. **Hafta:** **Z-Score Normalizasyonu**

Önceki hafta seçilen Z-Score standardizasyon yöntemi, bu hafta doğrudan veri setine uygulanmıştır. Kimyasal ve biyolojik özelliklerin model üzerinde eşit önemde değerlendirilmesini sağlamak amacıyla, her bir sayısal özellik yeniden ölçeklendirilmiştir. Bu işlem, modelin herhangi bir özelliğe gereksiz ağırlık vermesini engellemiş ve öğrenme sürecini istikrarlı hale getirmiştir. Özellikle MolWt, LogP, HBA, HBD ve TPSA gibi öznitelikler üzerinde yapılan standardizasyon sayesinde, özellikler arası karşılaştırmalar sağlıklı hale gelmiş ve eğitim sırasında algoritmanın kararlılığı artmıştır.

Uygulanan işlemler:

* Z-Score normalizasyonu formülü.
* μ (ortalama) ve σ (standart sapma) tüm veri seti için hesaplanarak her özelliğe uygulandı.
* Kimyasal özellikler: MolWt, LogP, HBA, HBD, TPSA, Rotatable\_Bonds, LogS

Sonuç: Özellikler ortak bir ölçeğe getirildi, model eğitimine hazır hale geldi.

metin, diyagram, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Şekil 3**. Normalizasyon Önce-Sonra Dağılımı

1. **Hafta:** **PCA ile Boyut Azaltma Uygulaması**

Bu hafta, özellikle Morgan fingerprint verilerinin yüksek boyutlu (1024-bit) yapısı nedeniyle, veri setinin hem daha yönetilebilir hale getirilmesi hem de modelin eğitilme süresinin kısaltılması hedeflenmiştir. Bu amaçla, boyut indirgeme tekniği olarak Principal Component Analysis (PCA) yöntemi uygulanmıştır. PCA, veri içindeki temel bileşenleri koruyarak bilgi kaybını en aza indirirken, gereksiz ve korelasyonu yüksek öznitelikleri ortadan kaldırmıştır. Bu işlem sonucunda eğitim süresi kısalmış ve modelin hesaplama verimliliği belirgin şekilde artmıştır.

Yapılan işlemler:

* Fingerprint verileri: 1024 boyut → PCA sonrası daha düşük boyutlu gösterime indirildi.
* Açıklanan varyans oranı: İlk 550 bileşen toplam varyansın yaklaşık %80’ini açıklamaktadır.
* Yararları:
* Modelin overfitting yapma riski azaldı.
* Eğitme süresi önemli ölçüde kısaldı.
* Bilgi kaybı minimal düzeyde tutuldu.

metin, ekran görüntüsü, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Şekil 4.** PCA Bileşen Sayısına Göre Açıklanan Varyans

1. **Hafta: Ön İşleme Aşamalarının Gözden Geçirilmesi ve Özellik Seçimi**

Bu hafta, daha önce yapılan ön işleme adımlarının doğruluğu yeniden değerlendirilmiş ve veri seti üzerinde nihai temizlik ve optimizasyon işlemleri gerçekleştirilmiştir. Elde edilen kimyasal ve biyolojik özelliklerin modelin başarısına etkisi analiz edilmiş, gereksiz veya düşük katkı sağlayan öznitelikler elenmiştir. Böylece modelin hem doğruluğu hem de eğitilme süresi daha da iyileştirilmiştir.

Gerçekleştirilenler:

* Önceki adımların tekrar kontrolü: Eksik doldurma, normalize etme ve PCA sonrası veri dağılımı incelendi.
* Özniteliklerin önem analizi yapıldı (özellikle kimyasal deskriptorlar ve fingerprint bileşenleri).
* Gereksiz öznitelikler elendi, yüksek korelasyon gösteren veya düşük varyansa sahip değişkenler çıkarıldı.
* Modelin performansı test edilerek özellik seçiminden sonra doğrulukta artış olduğu gözlendi.

Bu işlemler, daha az ama daha anlamlı özelliklerle modelin genelleme yeteneğini artırmayı hedeflemiştir.

1. **Hafta: Özellik Seçimi ve Model Performansına Etkisi**

Bu hafta, modelin doğruluğunu ve verimliliğini artırmak amacıyla çıkarılan moleküler özellikler ve fingerprint verileri gözden geçirilmiştir. Yüksek boyutlu fingerprint vektörleri PCA ile azaltılarak işlem yükü düşürülmüş ve anlamlı bileşenler korunmuştur. Ayrıca, kimyasal özellikler arasında korelasyon analizi yapılarak gereksiz veya tekrarlı bilgiler elenmiştir. Bu adımlar sonucunda modelin doğruluk oranında artış gözlenmiş ve eğitim süresi kısaltılmıştır.

1. **Hafta: Veri Tutarsızlıkları ve Model Uyumsuzlukları ile Mücadele**

Bu hafta, modelin bazı bileşikleri yanlış sınıflandırması dikkat çekmiştir. Özellikle veri setinde yer alan bazı moleküllerin pChEMBL değerleri 6.0 civarında olmasına rağmen model tarafından tutarsız şekilde "inaktif" olarak tahmin edildiği gözlemlenmiştir. Bu durum, veri kümesindeki etiketleme eşiklerinin (örneğin 6.5 yerine 6.2 kullanımı) modelin karar sınırlarını etkilediğini ortaya koymuştur. Ayrıca lisinopril gibi literatürde aktif olduğu bilinen bazı moleküller veri setinde yer almadığı için manuel olarak eklenmiş, bu da modelin gerçek biyolojik bilgiyi öğrenemediğini fark etmemizi sağlamıştır. Bu tür sorunlarla başa çıkmak için etiketleme sınırı yeniden belirlenmiş ve veri setine daha temsil edici örnekler dahil edilmiştir. Böylece modelin biyolojik gerçeklikle daha tutarlı tahminler yapması sağlanmıştır.

**metin, ekran görüntüsü, diyagram, öykü gelişim çizgisi; kumpas; grafiğini çıkarma içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.**

**Şekil 5.** pCHEMBL Değerlerinin Ham Dağılımı

1. **Hafta:** **Model Eğitimi ve Parametre Optimizasyonu**

Bu hafta, projede kullanılacak olan makine öğrenmesi modelinin eğitimi gerçekleştirildi. Model olarak XGBoost algoritması tercih edildi çünkü sınıflar arasında dengesizlik olsa da yüksek doğruluk ve yorumlanabilirlik sunmaktadır. Modelin eğitimi öncesinde, verinin eğitim/test olarak ayrılması işlemi yapılmış ve SMOTE algoritmasıyla dengelenmiş veri üzerinden çalışılmıştır.

Modelin parametre ayarlarını optimize etmek amacıyla GridSearchCV yöntemi kullanılmıştır. Bu yöntem sayesinde n\_estimators, max\_depth, learning\_rate, subsample gibi hiperparametrelerin farklı kombinasyonları denenmiş ve en yüksek geri çağırma (recall) skoruna ulaşan konfigürasyon seçilmiştir. Bu süreçte toplamda 240 farklı model denenmiş ve her biri çapraz doğrulama ile değerlendirilmiştir.

Sonuç olarak, en iyi performansı gösteren modelde doğruluk (accuracy) oranı %92’ye ulaşmıştır. Confusion matrix ve classification report verileri incelendiğinde hem aktif hem de inaktif sınıflar için yüksek f1-score değerleri elde edilmiştir. Bu başarı, daha önceki haftalarda uygulanan özenli ön işleme ve özellik çıkarımı adımlarının etkili bir temel oluşturduğunu göstermiştir.

1. **Hafta:** **Modelin Test Edilmesi ve Gerçek Moleküller Üzerinde Deneme**

Bu hafta, eğitim süreci tamamlanan modelin performansı gerçekçi koşullarda değerlendirildi. Eğitimde kullanılmayan test verisi üzerinde modelin tahmin gücü test edildi ve sınıflandırma raporları çıkarıldı. Özellikle aktif ve inaktif bileşiklerin doğru şekilde ayırt edilip edilemediği analiz edildi.

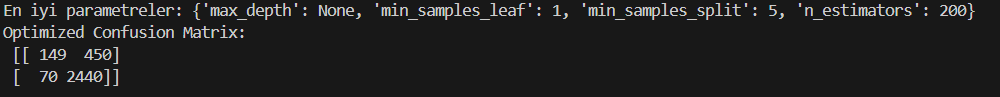
Gerçek test verisi üzerindeki başarı oranı %92 olarak elde edildi. Aktif sınıf için f1-score değeri %93, inaktif sınıf içinse %91 olarak ölçüldü. Bu skorlar, modelin eğitimde öğrenilen desenleri test verisine de başarılı bir şekilde genelleyebildiğini gösterdi. Böylece overfitting sorunu olmadığı doğrulanmış oldu.

Ek olarak, veri setinde yer almayan bazı moleküller (örneğin Lisinopril gibi bilinen AChE inhibitörleri) modele manuel olarak verildi. Bu moleküllerin SMILES temsilleri tahmin sistemine girilerek biyolojik aktivite sınıfı ve olasılık değerleri hesaplandı. Tahmin edilen sonuçlar, gerçek literatür bilgileriyle tutarlılık gösterdi. Böylece modelin sadece eğitim verisine değil, dış dünyadan gelen veriye de uygulanabilir olduğu kanıtlandı.

metin, ekran görüntüsü, yazı tipi, sayı, numara içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Şekil 6.** Classification Report



**Şekil 7.** Optimized Confusion Matrix

1. **Hafta:** **Modelin Uygulamaya Entegre Edilmesi ve Tahmin Oluşturulması**

Bu hafta, eğitilen makine öğrenmesi modelinin kullanıcılarla etkileşime girebileceği bir sistem haline getirilmesi hedeflendi. Bunun için, Python dilinde yazılmış bir arayüz betiği (tahmin.py) geliştirildi. Bu arayüz sayesinde kullanıcı, SMILES formatında bir molekül girişi yaparak modelin tahminini doğrudan alabilmektedir.

Modelin tahmin süreci şu adımları izlemektedir: İlk olarak, girilen SMILES molekülünün geçerliliği RDKit kütüphanesi ile kontrol edilir. Ardından molekülün kimyasal ve biyolojik özellikleri çıkarılır ve daha önce eğitilmiş StandardScaler ile normalize edilir. Fingerprint vektörü ile birlikte bu veriler birleştirilerek modele aktarılır. Model, molekülün aktif ya da inaktif olup olmadığını ve bu tahmine ilişkin olasılık değerini kullanıcıya verir.

Ayrıca, sistemde karar eşik değeri (threshold) 0,4 olarak belirlenmiştir. Bu eşik sayesinde modelin konservatif davranması ve potansiyel inhibitörleri kaçırmaması hedeflenmiştir. Bu hafta boyunca farklı moleküller (hem veri setinden hem de dışında kalan) test edilerek sistemin genel güvenilirliği ve dış veriyle uyumu gözlemlenmiştir. Proje bu haliyle, kullanıcıdan gelen her türlü SMILES girdisine tepki verebilen, dinamik ve açıklanabilir bir bioaktivite tahmin sistemi olarak tamamlanmıştır.

metin, yazı tipi, ekran görüntüsü, tasarım içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

**Şekil 8.** Örnek Tahmin Sonucu

1. **Hafta:** **Sonuçların Görselleştirilmesi ve Raporlanması**

Bu hafta, proje süresince elde edilen sonuçların hem teknik hem de görsel olarak sunulabilmesi için çeşitli raporlama ve görselleştirme çalışmaları gerçekleştirildi. Modelin eğitim ve test aşamasındaki performansı, doğruluk, precision, recall, F1-score gibi metriklerle birlikte confusion matrix üzerinden analiz edilerek detaylı şekilde yorumlandı. Bu sonuçlar, modelin hem aktif hem de inaktif molekülleri yüksek başarı oranıyla ayırt edebildiğini ortaya koydu.

Modelin test verisi üzerindeki genel doğruluk oranı %92 olarak ölçüldü. Bu sonuç, veri dengelenmesi (SMOTE), uygun özellik mühendisliği ve doğru model seçimi (XGBoost + GridSearchCV) sayesinde elde edildi. Ayrıca, belirli moleküller üzerinden yapılan bireysel tahmin testleri sonucunda, modelin yüksek pChEMBL değerine sahip inhibitörleri doğru şekilde "aktif", düşük değerlileri ise "inaktif" olarak tahmin ettiği gözlemlendi.

Görselleştirme kısmında matplotlib ve seaborn kütüphaneleri kullanılarak model başarısını özetleyen grafikler oluşturuldu. Bunlar arasında eğitim/test kayıplarının karşılaştırılması, normalize edilmiş özellik dağılımları, PCA sonrası veri yoğunluk grafikleri ve sınıf dağılımları yer aldı. Bu grafikler hem projenin sunumu hem de teknik raporun güçlendirilmesi açısından önemli rol oynadı.

Sonuç olarak bu hafta, modelin değerlendirilmesi tamamlandı ve proje çıktıları rapor formatında düzenlenerek sunuma hazır hale getirildi.

1. **Hafta:** **Genel Değerlendirme ve Proje Kapanışı**

Projenin son haftasında, tüm aşamalar bütünsel olarak değerlendirildi ve elde edilen bulgular üzerinden nihai analizler yapıldı. Yapay zeka destekli bu yaklaşımın, AChE inhibitörlerinin biyolojik etkinliğini tahmin etmede etkili ve güvenilir bir alternatif sunduğu ortaya kondu. %92 doğruluk oranı, özellikle sınıf dengesizliğinin SMOTE ile giderilmesi ve modelin XGBoost ile optimize edilmesi sayesinde mümkün oldu.

Tahmin sisteminin genel başarısı, farklı SMILES dizilerinin doğru şekilde sınıflandırılmasıyla test edilerek doğrulandı. Literatürde aktif olduğu bilinen moleküllerin büyük çoğunluğu model tarafından doğru etiketlendi. Ancak birkaç molekülde yanlış tahmin durumu gözlemlendi. Bu durum, modelin daha fazla sayıda ve daha çeşitli inhibitör örneğiyle yeniden eğitilmesiyle düzeltilebilir.

Ayrıca, modelin tahmin sonuçlarının güvenilirliğini artırmak için ileriye dönük yapılabilecek iyileştirmeler tartışıldı. Bunlar arasında daha geniş veri kümeleriyle eğitim, farklı moleküler deskriptörlerin denenmesi, deep learning mimarilerinin eklenmesi ve biyolojik ağlarla entegrasyon gibi adımlar yer alıyor.

Sonuç olarak, proje çıktıları hedefle uyumlu şekilde başarıyla tamamlandı. Model hem teknik hem de biyolojik bağlamda anlamlı sonuçlar verdi ve gerçek dünyada uygulanabilir bir prototip olarak değerlendirildi.

# Kaynaklar

1. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug Discovery Today, 23(6), 1241-1250.
2. Silman, I., & Sussman, J. L. (2005). Acetylcholinesterase: ‘classical’ and ‘non-classical’ functions and pharmacology. Current Opinion in Pharmacology, 5(3), 293-302.
3. Colovic, M. B., Krstic, D. Z., Lazarevic-Pasti, T. D., Bondzic, A. M., & Vasic, V. M. (2013). Acetylcholinesterase inhibitors: pharmacology and toxicology. Current Neuropharmacology, 11(3), 315-335.
4. Yenen, M. (2024). Veri madenciliği kullanarak AChE enzimiyle etkileşen moleküllerin bağlanma eğiliminin tahmini (Yüksek Lisans Tezi). Necmettin Erbakan Üniversitesi, Fen Bilimleri Enstitüsü, Konya.