



4. DÖNEM 2. GRUP VERİ ANALİTİĞİ PROJE RAPORU

PROTEİN – İLAÇ BAĞLANMA AFİNİTESİ TAHMİNİ

Hazırlayan: Beyza Fatıma Çekerekli

1. Giriş.....	2
1.1 İlgili Çalışmalar	2
2. Veri Kaynakları ve Yöntem	3
3. Veri Temizleme ve Hazırlama	3
4. Keşifsel Veri Analizi.....	4
4.2 LogP ve Moleküler Ağırlık ile pKi İlişkisi.....	5
4.3 Protein Bazında pKi Dağılımı	6
4.4 Sayısal Değişkenler Arası Korelasyon	7
4.5 Değişkenler Arası İkili İlişkiler	8
5. İstatistiksel Hipotez Testleri.....	9
6. Makine Öğrenmesi.....	10
7. Dashboard ve Genel Sonuç	12
8. Kısıtlamalar ve Öneriler	13
Kaynakça.....	14

1. Giriş

İlaç geliştirme sürecinde bir aday molekülün hedef proteine ne kadar güçlü bağlandığının belirlenmesi, klasik olarak deneysel yöntemlerle (in vitro tayinler) yapılır. Bu yöntemler güvenilir

olmakla birlikte zaman ve maliyet açısından sınırlıdır. Bu proje, halka açık bağlanma afinitesi veritabanlarından elde edilen molekül ve protein bilgilerini kullanarak, bir molekülün yalnızca yapısal özelliklerinden yola çıkarak bağlanma afinitesini (pK_i) tahmin edip edemeyeceğimizi incelemektedir.

Çalışma kapsamı kinaz ailesi proteinlerle sınırlı tutulmuştur; bu seçim, veri hacmini yönetilebilir kılmak ve biyolojik olarak tutarlı bir alt küme üzerinde çalışmak amacıyla yapılmıştır. Proje; veri toplama ve temizleme, keşifsel veri analizi, istatistiksel hipotez testleri, makine öğrenmesi ile tahmin modellemesi ve bulguların bir özet panelinde sunulması aşamalarından oluşmaktadır.

1.1 İlgili Çalışmalar

Kullanılan üç veri kaynağı da alan yazınında yerleşik, sık atıf alan kaynaklardır. BindingDB, protein-ligand kompleksleri için deneysel olarak ölçülmüş binlerce bağlanma afinitesi değerini bir araya getiren, literatürden derlenen bir veritabanıdır [3]. Davis veri seti, 72 kinaz inhibitörünün insan kinazomunun %80'inden fazlasını kapsayan 442 kinaz proteiniyle etkileşimini biyokimyasal bir deney paneliyle sistematik olarak ölçmüştür [1]. KIBA veri seti ise K_i , K_d ve IC_{50} gibi farklı ölçüm türlerini "KIBA score" adı verilen tek bir bütünleşik puana indirgeyen model tabanlı bir yaklaşım sunar [6]. Bu çalışmada farklı bir ölçek kullanması nedeniyle nihai veri setine dahil edilmemiştir.

Moleküllerin fizikokimyasal tanımlayıcılar üzerinden değerlendirilmesi yaklaşımı, ilaç keşfinde uzun süredir kullanılan Lipinski'nin "Beşler Kuralı" geleneğine dayanır; bu kural, biyoyararlanımı yüksek bileşiklerin moleküler ağırlık, $LogP$, hidrojen bağı verici/alıcı sayısı gibi belirli sınırlar içinde kaldığını öne sürer [2]. Bu proje bu tanımlayıcıların bir kısmını (MW , $LogP$, HBD , HBA) ve ek olarak $TPSA$ ile döndürülebilir bağ sayısını kullanmaktadır. Random Forest problemlerinde etkili bir yöntem olduğu, bileşik sınıflandırma ve biyolojik aktivite tahmininde erken dönemde gösterilmiştir [5].

Derin öğrenme tabanlı yaklaşımlar da bu alanda önemli bir yer tutmaktadır; örneğin DeepDTA, SMILES ve protein dizisi temsillerini doğrudan girdi olarak kullanan bir evrişimli sinir ağı mimarisiyle bağlanma afinitesini tahmin etmekte ve bu çalışmada da Davis veri setindeki $pK_d=5.0$ tabanı gibi veri özelliklerinin yorumlanmasında referans alınmıştır. [4].

2. Veri Kaynakları ve Yöntem

Proje üç farklı veritabanından beslenmiştir:

- **BindingDB** — Ki, IC50 ve Kd değerlerini (nM cinsinden) içeren geniş kapsamlı bir bağlanma afinitesi veritabanı. Hedef adında "kinase" ifadesi geçen kayıtlarla sınırlandırılmıştır.
- **Davis** — Kinaz inhibitörleri için pKd (log ölçekli) değerleri doğrudan içeren, daha küçük ve odaklı bir veri seti.
- **KIBA** — Ki, Kd ve IC50 değerlerini birleştiren kendine özgü bir puanlama sistemi (KIBA score) kullanır; bu ölçek diğer ikisiyle doğrudan karşılaştırılabilir olmadığından, veri toplanmış ancak nihai birleşik veri setine dahil edilmemiştir.

Genel iş akışı şu şekildedir: verilerin yüklenmesi → ortak bir pKi ölçeğine dönüştürülmesi → RDKit kütüphanesi ile moleküler tanımlayıcıların hesaplanması → veri setlerinin birleştirilmesi ve temizlenmesi → keşifsel analiz → istatistiksel testler → makine öğrenmesi modellemesi.

3. Veri Temizleme ve Hazırlama

BindingDB dosyası yaklaşık 9 GB boyutunda olduğundan, tamamının belleğe yüklenmesi yerine parça parça ve yalnızca ihtiyaç duyulan sütunlar (Ligand SMILES, Target Name, Ki/IC50/Kd) okunarak, "kinase" ifadesi geçmeyen satırlar anında elenmiştir. Bu filtreleme sonucunda 721.649 satırlık bir alt küme elde edilmiştir.

Ki, IC50 ve Kd sütunları pandas melt() işlemiyle tek bir "affinity_value" sütununda birleştirilmiş (694.000 satır), değerlerdeki "<" ve ">" gibi sansür işaretleri sayısal dönüşüme izin vermek için temizlenmiş, ardından tüm değerler aşağıdaki formülle pKi ölçeğine çevrilmiştir:

$$pKi = -\log_{10}(\text{değer_nM} / 1.000.000.000)$$

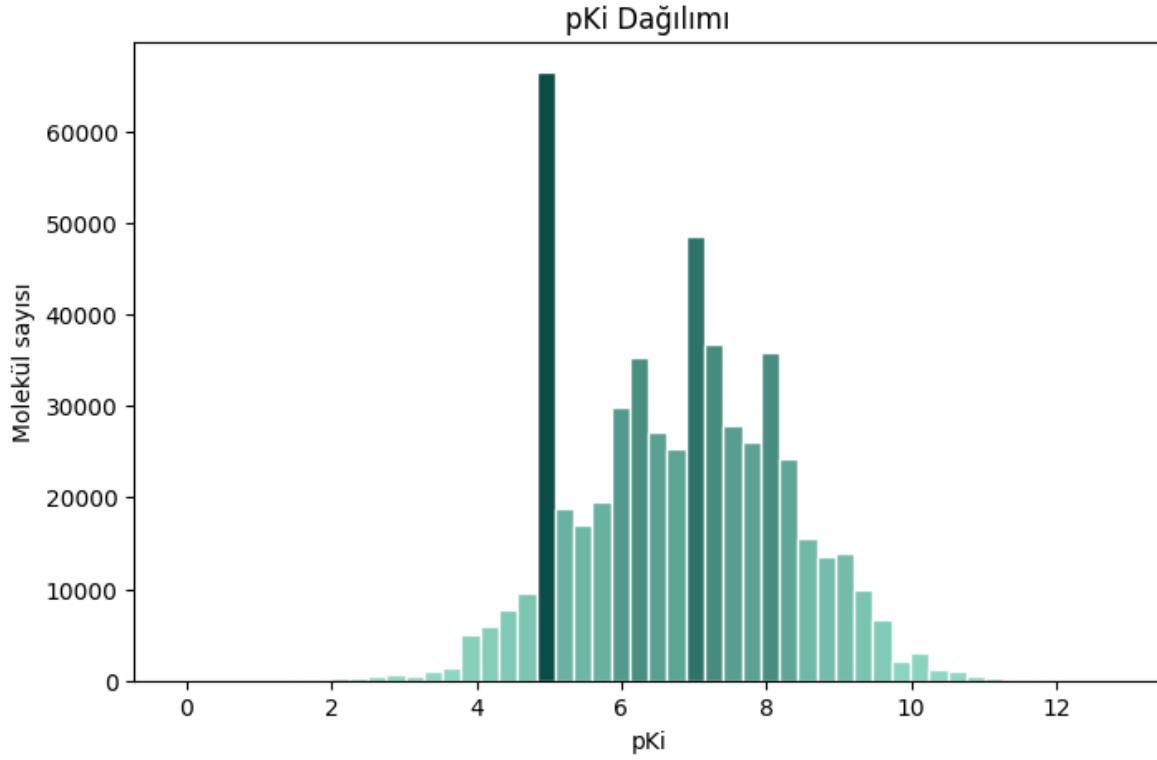
Davis veri setindeki afinite değerleri zaten log ölçekli (pKd) olduğundan doğrudan pKi olarak kabul edilmiştir (30.056 satır). BindingDB ve Davis verileri birleştirildikten sonra (724.014 satır), 291.003 benzersiz molekül için RDKit ile altı tanımlayıcı hesaplanmıştır: Moleküler Ağırlık (MW), LogP, Topolojik Kutupsal Yüzey Alanı (TPSA), Hidrojen Bağ Verici Sayısı (HBD), Hidrojen Bağ Alıcı Sayısı (HBA) ve Döndürülebilir Bağ Sayısı (RotBonds). Geçersiz SMILES gösterimine sahip moleküller (RDKit tarafından ayrıştırılamayanlar) veri setinden çıkarılmıştır.

Aynı molekül-protein çiftinin birden fazla kez ölçülmüş olabileceği göz önünde bulundurularak, tekrarlanan kayıtların pKi değerlerinin ortalaması alınmıştır. Bu adım sonunda satır sayısı 723.736'dan 539.218'e düşmüştür. Nihai **temizlenmiş veri seti** 539.218 satır ve 10 sütundan oluşmaktadır: SMILES, protein, source, pKi, MW, LogP, TPSA, HBD, HBA, RotBonds.

4. Keşifsel Veri Analizi

Temizlenmiş veri seti üzerinde altı görsel analiz gerçekleştirilmiştir.

4.1 pKi Dağılımı

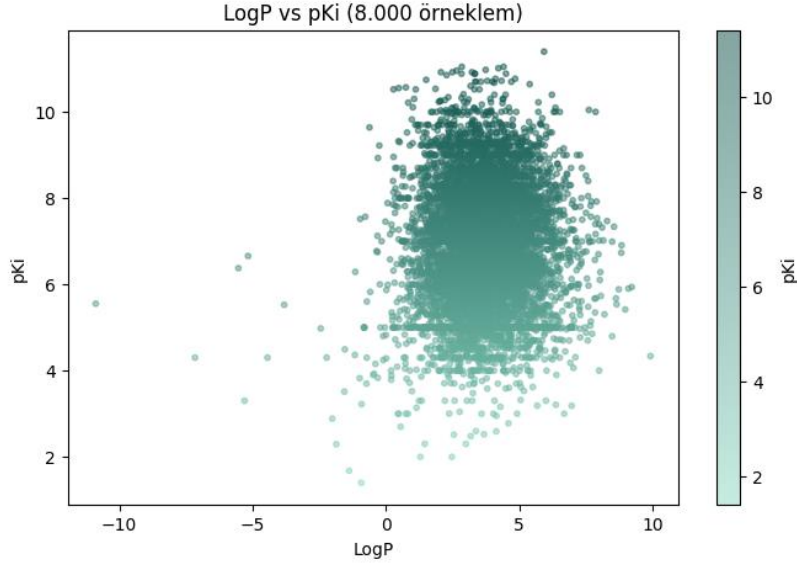


Şekil 1. pKi değerlerinin dağılımı (539.218 molekül-protein çifti)

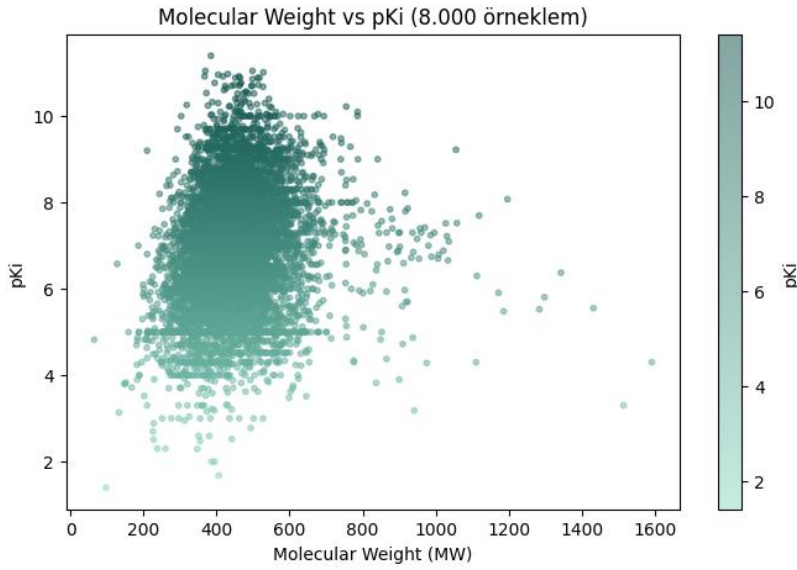
Dağılımda $pK_i=5.0$ ve $pK_i=7.0$ civarında belirgin yığılmalar dikkat çekmektedir. $pK_i=5.0$ 'te 58.569 kayıt (BindingDB: 40.226, Davis: 18.343) bulunmaktadır. Bu yığılma, veri toplama sürecinden kaynaklanan bir yapaylıktır: Davis veri setinde, ölçülebilir eşiğin altında kalan (etkisiz kabul edilen) bağlanmalar için $pK_d=5.0$ standart bir taban değeri olarak atanmaktadır [4]. $pK_i=7.0$ 'de ise 24.179 kayıt (BindingDB: 24.144, Davis: 35) bulunmakta olup bu yığılma büyük ölçüde BindingDB'de $K_i=100$ nM etrafındaki kümelenmeden kaynaklanmaktadır.

Bu bulgular ham veriyi okumadan varsayımda bulunmak yerine, kaynak bazında `value_counts()` ile doğrulanmış ve ilgili literatürle karşılaştırılmıştır.

4.2 LogP ve Moleküler Ağırlık ile pKi İlişkisi



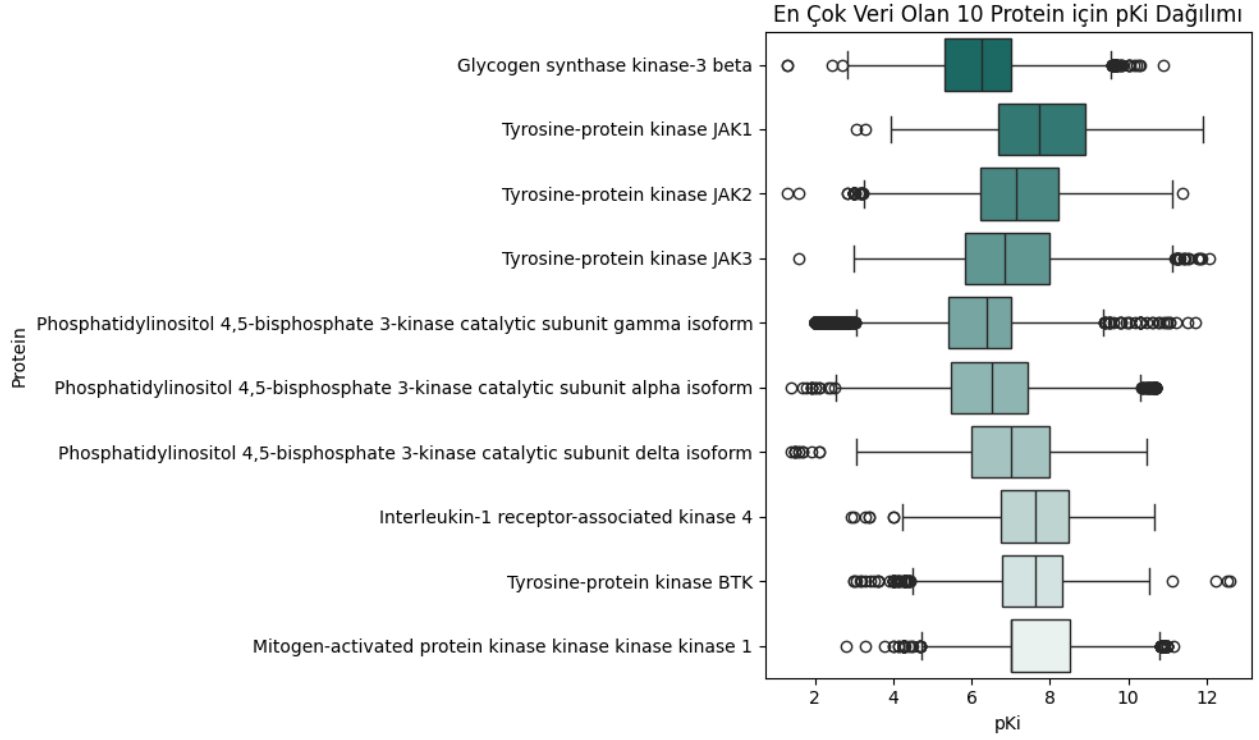
Şekil 2. LogP vs pKi (8.000 örneklem)



Şekil 3. Moleküler Ağırlık vs pKi (8.000 örneklem)

Her iki grafikte de belirgin bir doğrusal örüntü gözlenmemektedir; noktalar geniş bir bulut şeklinde dağılmıştır. Bu, tek bir fizikokimyasal özelliğin pKi'yi tek başına açıklamaya yetmediğine işaret etmektedir. Bu gözlem, ilerleyen makine öğrenmesi bölümündeki düşük R^2 değerleriyle de tutarlıdır.

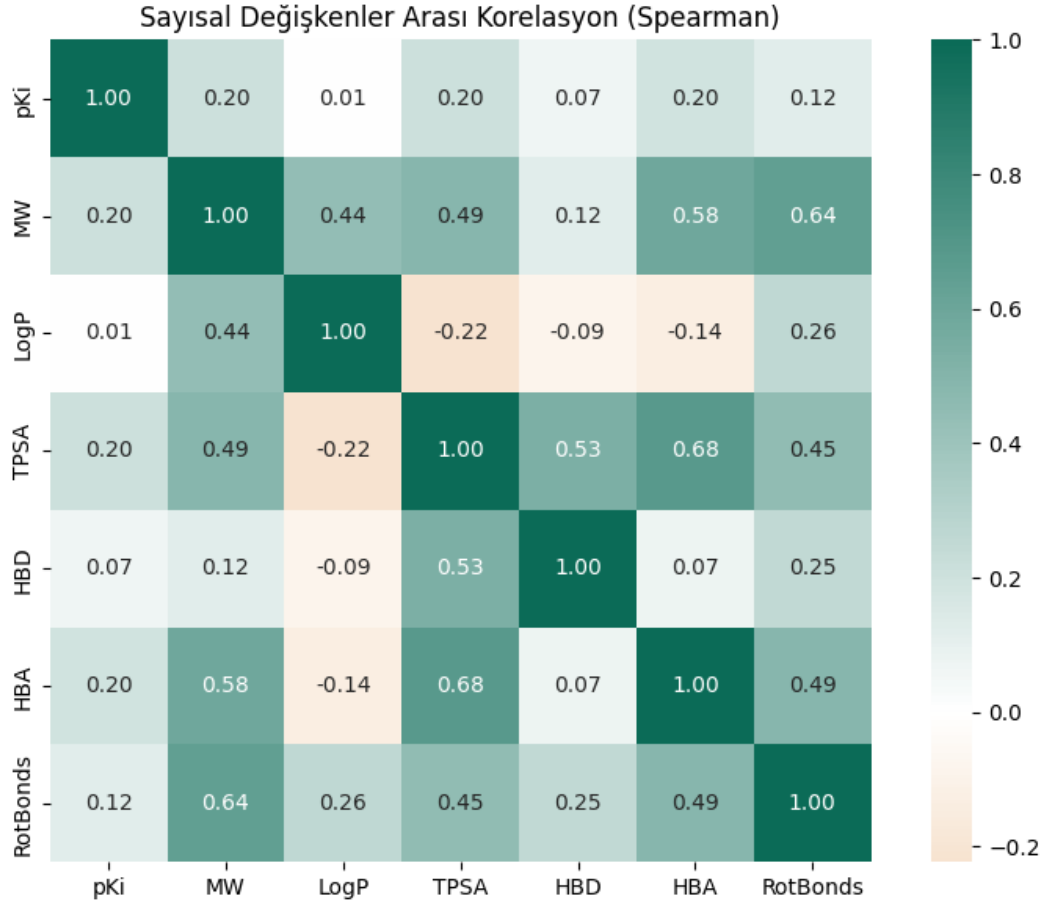
4.3 Protein Bazında pKi Dağılımı



Şekil 4. En çok veri içeren 10 protein için pKi dağılımı (kutu grafiği)

Farklı protein hedefleri arasında pKi medyan ve yayılım değerlerinde gözle görülür farklılıklar bulunmaktadır. Bu gözlem, ileride Bölüm 5'te Kruskal-Wallis testiyle istatistiksel olarak da doğrulanmıştır.

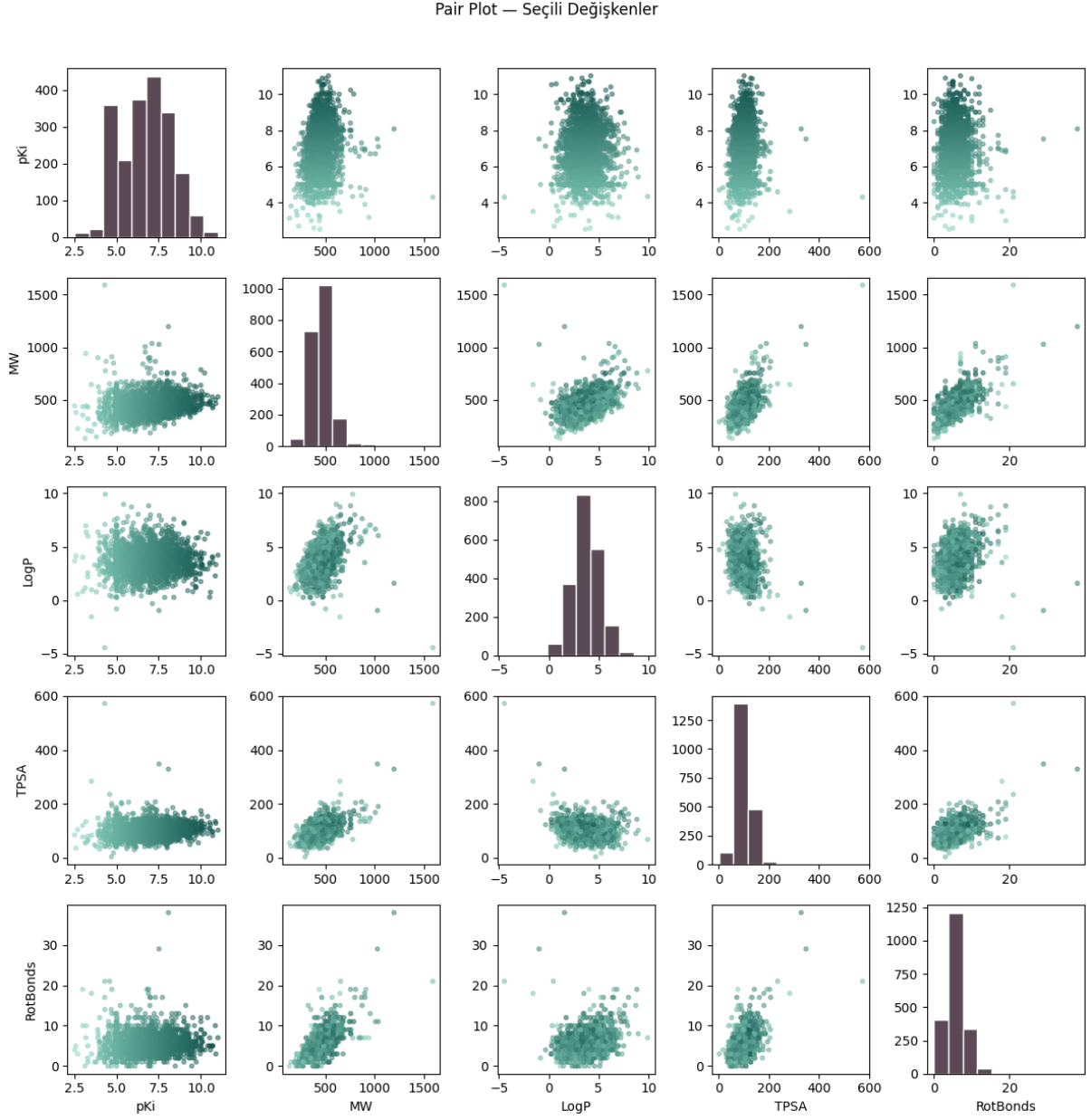
4.4 Sayısal Değişkenler Arası Korelasyon



Şekil 5. Spearman korelasyon matrisi

Korelasyon matrisine göre, pKi ile **en güçlü ilişkiyi** gösteren değişkenler MW ($\rho=0.20$), TPSA ($\rho=0.20$) ve HBA ($\rho=0.20$) olup, LogP ile ilişki neredeyse sıfıra yakındır ($\rho=0.01$). Tanımlayıcıların kendi aralarında ise daha güçlü ilişkiler mevcuttur; örneğin TPSA-HBA ($\rho=0.68$) ve MW-RotBonds ($\rho=0.64$). Bu durum çoklu doğrusal bağlantı açısından dikkate alınmalıdır.

4.5 Değişkenler Arası İkili İlişkiler



Şekil 6. Pair plot — pKi, MW, LogP, TPSA, RotBonds (2.000 örneklem)

Pair plot, ikili değişken kombinasyonlarında da güçlü doğrusal örüntülerin bulunmadığını, ancak pKi'nin diğer değişkenlerle karşılaştırıldığında görece dar bir aralıkta yoğunlaştığını göstermektedir.

5. İstatistiksel Hipotez Testleri

Testlere başlamadan önce pKi değişkeninin normal dağılıp dağılmadığı **Shapiro-Wilk testiyle** (5.000 kişilik bir örneklem üzerinde, çünkü test büyük örneklerde aşırı hassaslaşmaktadır) kontrol edilmiştir: $W=0.9861$, $p \approx 0.000 \rightarrow$ pKi normal dağılmamaktadır. Bu nedenle, tüm hipotez testlerinde parametrik olmayan yöntemler (Spearman korelasyonu ve Kruskal-Wallis testi) tercih edilmiştir.

Hipotez	Değişken	Korelasyon / İstatistik	p-değeri	Sonuç
H1	MW vs pKi	$\rho = 0.204$	≈ 0.000	Anlamlı
H2	LogP vs pKi	$\rho = 0.010$	$2.37e-12$	Anlamlı (etkisi ihmal edilebilir)
H3	TPSA vs pKi	$\rho = 0.205$	≈ 0.000	Anlamlı (yön beklenenin tersi)
H4	HBA vs pKi	$\rho = 0.196$	≈ 0.000	Anlamlı
H5	Protein grubu vs pKi	$H = 12322.09$	≈ 0.000	Anlamlı

H3 için başlangıçta TPSA arttıkça pKi'nin azalması (negatif ilişki) beklenmiştir; ancak test sonucu zayıf da olsa pozitif bir ilişki ($\rho=0.205$) göstermiştir. Sonucunda hipotez yön bakımından desteklenmemiştir. Bu, verinin önceden var sayılmadan, gerçek sonuçlara göre yorumlanması gerektiğinin bir örneğidir.

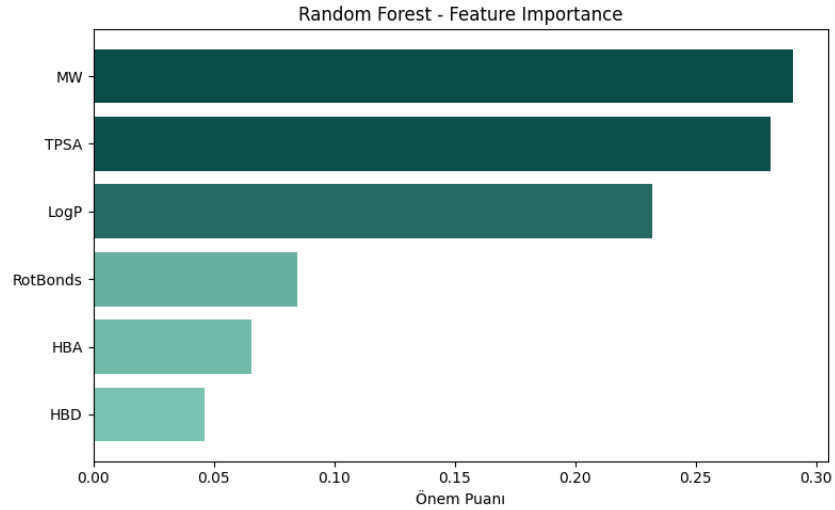
Tüm testlerde p-değerinin 0.05 eşiğinin çok altında çıkması "**istatistiksel anlamlılık**" ifade eder; ancak korelasyon katsayılarının 0.01–0.21 aralığında kalması, bu ilişkilerin pratikte zayıf olduğunu göstermektedir. 539.218 gözlemlik bir örneklemde, çok küçük etkiler bile $p < 0.05$ sonucu verebilmektedir. Bu nedenle p-değeri tek başına değil, korelasyon katsayısı ile birlikte değerlendirilmiştir.

6. Makine Öğrenmesi

Modelleme için özellik kümesi olarak MW, LogP, TPSA, HBD, HBA ve RotBonds; hedef değişken olarak pKi kullanılmıştır. Veri seti %80 eğitim (431.374 satır) ve %20 test (107.844 satır) olacak şekilde ayrılmış, doğrusal modeller için StandardScaler ile ölçeklendirme uygulanmıştır. Altı model eğitilmiştir: Linear Regression, Ridge, Lasso, Decision Tree, Random Forest ve Gradient Boosting. XGBoost, macOS ortamında eksik bir sistem bağımlılığı (libomp) nedeniyle çalıştırılamadığından, benzer bir topluluk (ensemble) yöntemi olan Gradient Boosting ile değiştirilmiştir.

Model	Test R ²	RMSE	MAE
Random Forest	0.5171	1.0063	0.7425
Decision Tree	0.3340	1.1818	0.8261
Gradient Boosting	0.1486	1.3362	1.0904
Linear Regression	0.0486	1.4125	1.1594
Ridge	0.0486	1.4125	1.1594
Lasso	0.0481	1.4129	1.1605

En iyi performansı Random Forest göstermiştir (Test R²=0.517), ancak eğitim setindeki **R² değeri (0.800)** ile karşılaştırıldığında bir miktar aşırı öğrenme (**overfitting**) olduğu görülmektedir. Decision Tree'de bu fark çok daha belirgindir (Eğitim R²=0.829, Test R²=0.334), yani model eğitim verisini ezberlemiş ancak yeni veriye iyi genelleymemiştir. Gradient Boosting'de ise tam tersi bir durum söz konusudur: hem eğitim (R²=0.151) hem test (R²=0.149) performansı düşüktür, bu da modelin yetersiz öğrendiğine (**underfitting**) işaret etmektedir. Linear Regression, Ridge ve Lasso birbirine çok yakın ve düşük sonuçlar vermiştir; bu da veride güçlü doğrusal bir örüntü bulunmadığını doğrulamaktadır.



Şekil 7. Random Forest özellik önem dereceleri

Özellik	Önem Puanı
MW	0.2905
TPSA	0.2811
LogP	0.2319

RotBonds	0.0845
HBA	0.0656
HBD	0.0463

Random Forest'a göre **en belirleyici üç özellik** MW, TPSA ve LogP'dir; bu üçü toplam önemin yaklaşık %80'ini oluşturmaktadır. Dikkat çekici bir nokta: LogP'nin Bölüm 5'teki basit Spearman korelasyonu neredeyse sıfırken ($\rho=0.01$), Random Forest'ta üçüncü en önemli özellik olarak çıkmasıdır. Bu, LogP'nin pKi ile tek başına doğrusal bir ilişkisi olmasa da, diğer özelliklerle etkileşim hâlinde tahmine katkı sağladığını göstermektedir. Korelasyon matrisi yalnızca ikili doğrusal ilişkileri yakalarken, Random Forest bu tür etkileşim etkilerini de görebilmektedir.

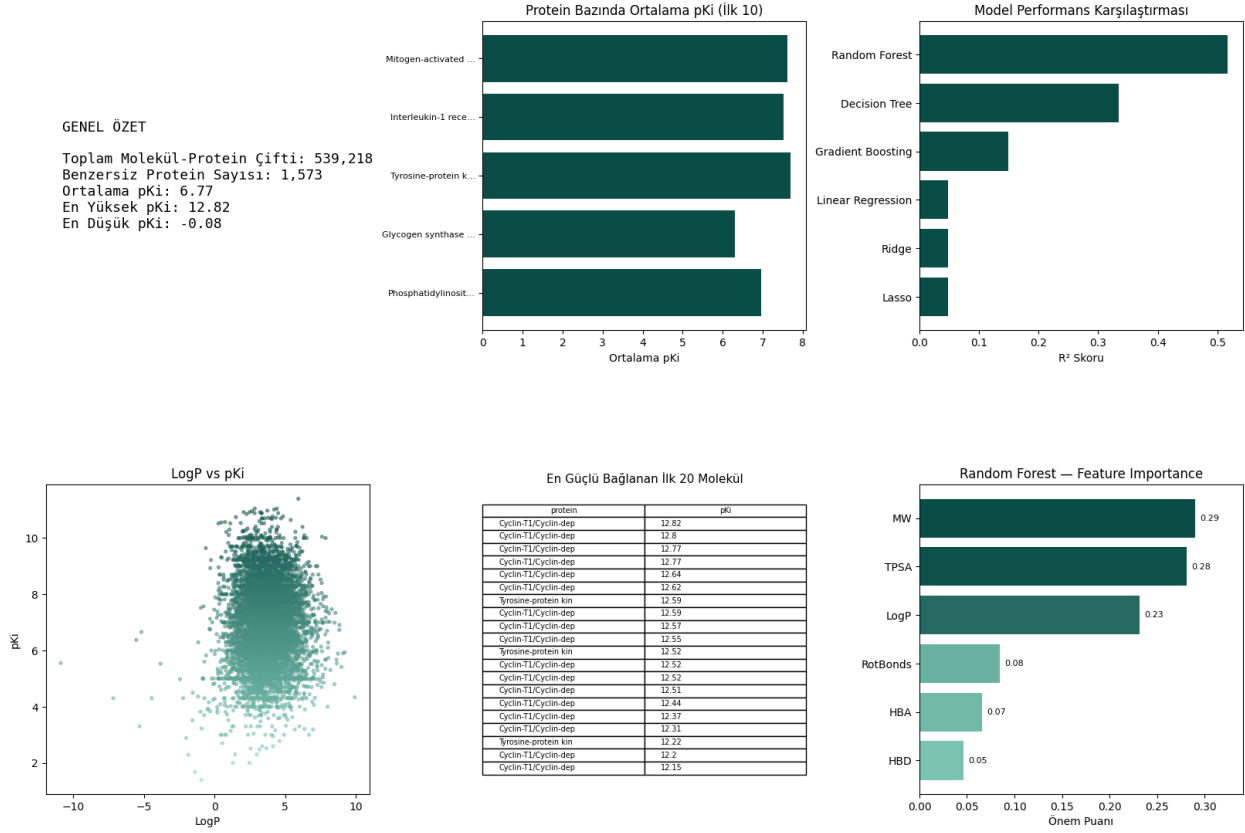
Genel olarak modellerin açıklama gücünün sınırlı kalmasının (en iyi modelde bile %52) en olası nedeni, kullanılan özelliklerin yalnızca ligand (molekül) tabanlı olmasıdır. Protein kimliği veya yapısı modele özellik olarak dahil edilmemiştir. Ancak bölüm 5'teki H5 testi, protein hedefinin pKi üzerinde çok güçlü bir etkisi olduğunu göstermektedir (Kruskal-Wallis istatistiği: 12322.09).

Aynı fizikokimyasal özelliklere sahip bir molekül, farklı proteinlere farklı güçte bağlanabilir; bu bilgi mevcut modellerde yakalanamamaktadır.

7. Dashboard ve Genel Sonuç

Projenin tüm bulguları tek bir özet panelde bir araya getirilmiştir:

Protein-ilaç Bağlanma Afinitesi — Proje Dashboard



Şekil 8. Proje özet dashboard'ı

Özetle, 539.218 molekül-protein çifti ve 1.573 benzersiz protein içeren bir veri seti üzerinde çalışılmış, ortalama pKi 6.77 olarak hesaplanmıştır. Keşifsel analiz, fizikokimyasal özelliklerin pKi ile zayıf ama istatistiksel olarak anlamlı ilişkiler taşıdığını; hipotez testleri, protein hedefinin pKi üzerinde güçlü bir etkisi olduğunu; makine öğrenmesi ise Random Forest'ın (Test $R^2=0.517$) bu özelliklerle en iyi tahmini sağladığını göstermiştir.

8. Kısıtlamalar ve Öneriler

- BindingDB verisi yalnızca hedef adında "kinase" ifadesi geçen kayıtlarla filtrelenmiştir; bulgular kinaz ailesiyle sınırlıdır, tüm protein evrenine genellenemez.
- K_i , IC_{50} ve K_d farklı deneysel ölçüm türleri olup biyokimyasal olarak tam eşdeğer değildir; bu çalışmada ortak bir pK_i ölçeğinde birleştirilmiştir.
- Değerlerdeki "<" ve ">" sansür işaretleri kaldırılıp değerler nokta tahmini olarak kullanılmıştır; bu, sınır değerlerindeki belirsizliği bir miktar gizlemektedir.
- Davis veri setindeki $pK_d=5.0$ tabanı, muhtemelen gerçek bir ölçüm değil, eşik altı bağlanmalar için atanan bir kongvansiyondur.
- KIBA veri seti toplanmış, ancak farklı bir ölçek kullandığından nihai birleşik veri setine dahil edilmemiştir.
- Makine öğrenmesi modelleri yalnızca ligand tabanlı 2D tanımlayıcıları kullanmış, protein kimliği veya yapısı modele dahil edilmemiştir — bu, model performansını sınırlayan en önemli etkidir.
- Ortam kısıtları nedeniyle XGBoost yerine Gradient Boosting kullanılmıştır.

Gelecek çalışmalar için öneriler: protein dizisi veya yapısına dayalı özelliklerin (örn. protein embedding'leri) modele eklenmesi, kinaz dışı protein ailelerine genişletilmesi ve sansürlü veri noktalarının ayrı bir kategori olarak ele alınması önerilmektedir.

Kaynakça

- [1] Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., Hocker, M., Treiber, D. K., & Zarrinkar, P. P. (2011). Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29(11), 1046–1051.
- [2] Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (1997). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23(1–3), 3–25.
- [3] Liu, T., Lin, Y., Wen, X., Jorissen, R. N., & Gilson, M. K. (2007). BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Research*, 35, D198–D201.
- [4] Öztürk, H., Özgür, A., & Ozkirimli, E. (2018). DeepDTA: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17), i821–i829.

[5] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of Chemical Information and Computer Sciences*, 43(6), 1947–1958.

[6] Tang, J., Szwajda, A., Shakyawar, S., Xu, T., Hintsanen, P., Wennerberg, K., & Aittokallio, T. (2014). Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54(3), 735–743.