Süleyman Demirel Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
Cilt 25, Sayı 2, 365-370, 2021

Süleyman Demirel University
Journal of Natural and Applied Sciences
Volume 25, Issue 2, 365-370, 2021

DOI: 10.19113/sdufenbed.867067

In-Silico Mutajenisite Tahmininde İstatistiksel Öğrenme Modeli

Enis GÜMÜŞTAŞ[1], Ayça ÇAKMAK PEHLİVANLI[*2]

[1,2]Mimar Sinan Güzel Sanatlar Üniversitesi, Fen Edebiyat Fakültesi, İstatistik Bölümü, 34380, İstanbul, Türkiye

(Alınış / Received: 23.01.2021, Kabul / Accepted: 02.03.2021, Online Yayınlanma / Published Online: 15.08.2021)

Anahtar Kelimeler
Sınıflama,
Topluluk Öğrenmesi,
XGBoost,
LightGBM,
Değişken Seçimi,
Toksisite

Özet: Toksisite testleri arasında, bir etken nedeniyle ortaya çıkabilecek genetik değişim (mutasyon) olarak tanımlanabilen mutajenisite önemli yer tutmaktadır. Bu çalışmada genel olarak mutajenisite belirleme sürecini iyileştirebilmek adına in-silico yaklaşım kapsamında istatistiksel öğrenme algoritmaları kullanılmıştır. Söz konusu yaklaşım deneyler ile elde edilen mutajenisite bilgisi içeren molekül setine uygulanmış ve dikkate değer sınıflama başarıları elde edilmiştir. Çalışmada kullanılmak üzere literatürde bulunan, moleküllerden oluşan Bursi ile Benchmark veri setleri birleştirilmiş ve Molecular Operating Environment (MOE) programı aracılığı ile moleküllerin özellikleri hesaplanmıştır. Hesaplama sonucunda 10835 gözleme ve 193 değişkene sahip veri seti üzerinde karar ağaçları algoritmaları uygulanarak grid arama yaklaşımı ile parametre seçimi gerçekleştirilmiştir. Elde edilen en iyi parametreler ile kurulan modeller sonucunda değişkenlerin seçimi mutajenisiteyi tahmin etmedeki önem düzeylerine göre yapılmış ve verinin boyutu en etkili 72 değişkene indirgenmiştir. Seçilen değişkenlerden oluşan yeni veriye farklı istatistiksel öğrenme algoritmaları uygulanmış ve içlerinden en iyi sonuç veren beş sınıflama algoritmasına karar verilmiştir. Parametre en iyilemesi ile model başarımları arttırılan bu algoritmalar kullanılarak yaklaşık %90 mutajenisiteyi doğru sınıflama oranları elde edilmiştir.

Statistical Learning Model for In-Silico Mutagenicity Prediction

Keywords
Classification,
Ensemble Learning,
XGBoost,
LightGBM,
Feature Selection,
Toxicity

Abstract: Among the toxicity tests, mutagenicity defined as a genetic change that can occur due to an agent, has an important place. In this study, statistical learning algorithms were used within the scope of in-silico approach in order to improve the mutagenicity determination process in general. This approach has been applied to the set of molecules containing mutagenicity information obtained by experiments and remarkable classification success were achieved. In order to use in this study, Bursi and Benchmark data sets consisting of molecules found in the literature were combined and the properties of molecules were calculated by means of the Molecular Operating Environment (MOE). As a result of the calculation, decision trees algorithms were applied on the data set with 10835 molecules and 193 variables and parameter selection was performed with grid search approach. The selection of variables was made according to their level of importance in predicting mutagenicity as a result of models established with the best parameters obtained, and the number of descriptors variables was reduced to the 72 most effective descriptor variables. Various statistical learning algorithms were applied to the reduced data set consisting of the selected variables, and five classification algorithms with the best results were decided. By the algorithms whose model performances were increased by means of parameter optimization, accurate prediction rates were obtained approximately 90% for mutagenicity classification.

1. Giriş

Gelişen teknolojilerle birlikte canlı organizmanın dışında yapılan in-vitro ve canlı organizma üzerinde yapılan in-vivo deneyler yerini laboratuvar deneylerine gereksinim duymadan bilgisayar ortamında geliştirilen istatistiksel ve hesaplamalı yöntemlere bırakmaya başlamıştır. Genel olarak in-silico adı verilen bu yöntemler in-vivo ve/ya in-vitro testlere geçmeden önce aday ilaç moleküllerine yönelik öngörme, önbilgi verebilme yetkinliğindedir.

*İlgili yazar: ayca.pehlivanli@msgsu.edu.tr

Doğru bir in silico yaklaşım, moleküle ait elde edilen bilginin laboratuvar deneylerine geçilip geçilmemesi konusunda yönlendirici olmasının yanında yapılacak testlerin tasarımında daha az deney hayvanı kullanılması, kullanılacak konsantrasyonun önceden belirlenebilmesi, zaman ve maliyetin azaltılabilmesi gibi avantajlar da sağlayabilir.

Günümüzde kimyasalların yasal düzenlemelerinde (ilaç molekülleri, gıda katkı maddeleri, kozmetik gibi) çeşitli toksisite testlerinden yararlanılır. Özellikle aday ilaç moleküllerinin klinik çalışmalarına devam edebilmesi için bir etken nedeniyle ortaya çıkabilecek genetik değişime (mutasyon) karşılık gelen mutajenik etkilerinin olmaması önkoşuldur. Toksisiteye yol açması nedeniyle mutajenisitenin önceden tespiti çok önemlidir. Bu tespitin yapılabilmesi için de çeşitli yöntemlerin yanı sıra istatistiksel öğrenme yaklaşımları çoklukla kullanılmaktadır.

In silico çalışmalar genel olarak kurala ve uzman bilgisine dayalı sistemler ve kantitatif yapı-aktivite ilişkisi (QSAR) olarak da bilinen istatistiksel yöntemlere dayalı yaklaşımlar olmak üzere iki grupta toplanabilir [1-3]. Özellikle 90'ların sonunda hız kazanan kurala ve uzman bilgisine dayalı sistemlere ilişkin erken dönem çalışmalarında kimyasal yapılar ile gözlemlenen toksik çıktılar arasındaki ilişkiler incelenmiş, çeşitli yazılımlar karşılaştırmalı olarak ortaya konmuştur [4]. Bu çalışmalara aleternatif olarak istatistiksel yöntemlere dayalı modellemeler ve yazılımlar özellikle moleküllerin fizikokimyasal özelliklerine dayalı olarak biyolojik aktivitelerini çeşitli makine öğrenmesi algoritmaları ile tahmin etmekte kullanılmıştır [5]. Mazzatorta ve ark. mutajenisite tahmini için bu iki yaklaşımı birleştiren hibrit bir in silico yaklaşım önermiş ve test verisi üzerinde %85 tahmin başarısı elde etmişlerdir [6]. 2000'li yılların başlarında bu alanda yapılan çalışmalarda in silico yöntem olarak özellikle destek vektör motroları yaygın olarak tercih edilmiştir. Zheng ve ark. kimyasal moleküllerin mutajenik olasılıklarının tahmininde %85 başarı elde ederken, Liao ve ark. yinelemeli bölünme ile seçtikleri değişkenlerden oluşan üç farklı veri setinden destek vektör motorları kullanarak %81.4, %87 ve %87.3 oranında sınıflama başarısına ulaşmışlardır [7-8].

2012 yılında Xu ve ark. tarafından mutajenisiste tahmininde %56'sı mutajen olan 7617 farklı bileşik içeren bir çalışma yapılmıştır. Çalışmada yaygın kullanılan beş farklı öğrenme yöntemi ile model oluşturmuş, oluşan modellerin performansı 831 farklı bileşik içeren harici veri seti ile sınanmış ve %90.4 ile %98 arası başarı elde edilmiştir [9] Son yıllarda bu alanda çeşitli çalışmalar yapılmıştır. Moorthy ve ark. veri madenciliği algoritmaları ile elde ettikleri modelleri sıra fark toplamı ile sıralamış ve rastgele orman algoritması ile %70 başarı elde etmiştir. 2017 yılında yapılan bu araştırmada 1481 molekülden oluşan veri setinin mutajenik ve kanserojen bilgileri kullanmıştır [10]. Yine aynı yıl Zhang ve arkadaşları yeni bir Naive Bayes yaklaşımı önermiş ve ilaç moleküllerinin tahmininde ekili olan mutajenisite bilgisi içeren veri seti üzerine uygulamıştır. Kullandığı farklı sınama veri setleri üzerinde %70.3 ve %90.9 düzeyinde başarı elde etmişlerdir [11]. Webb ve arkadaşları destek vektör makineleri ile rastgele orman algoritmalarını kullanarak mutajenisite verisi üzerinde 2014'te yaptıkları çalışmada %82 başarı elde etmişlerdir [12]. Seal ve arkadaşları tarafından üç farklı mutajenisite veri seti üzerine çeşitli sınıflama algoritmaları uygulanmış ve %79 ile % 85 arasında başarı elde edilmiştir [13].

Son yıllarda uygulamalı matematik, istatistik ve bilgisayar alanlarında yaşanan gelişmeler, kimya, biyoloji ve genetikte kullanılan karmaşık sistemlerin çözümü için yeni disiplinlerin doğuşuna neden olmuştur. Bu disiplinler sayesinde kimyasal ve biyolojik verilerden gerçek bilginin elde edilebilmesi veya saklı bilgilerin açığa çıkarılması sağlanabilmektedir. Bu kapsamda Ji ve arkadaşları tarafından 2019 yılında kombine ilaç tedavilerinin iyileştirme etkisini artırmak ve beraberinde gelen yan etkileri azaltmak amacı ile yapılan çalışmada ilaçlar sinerjist ve antagonist etkiler bakımından sınıflandırılmıştır. Çalışmada XGBoost algoritması, ilaçların beş özelliği temel alınarak uygulanmış ve in-silico yaklaşımın deneysel yaklaşımlara göre çok daha etkili olduğu gösterilmiştir [14].

Biyolojik moleküllerden elde edilen verilerin boyutu çok büyük olabilmektedir. Büyük boyutlu veriler ise daha yüksek hesaplama gücüne ve buna bağlı olarak artan hesaplama sürelerine ihtiyaç duymaktadır. Bunun yanı sıra gereksiz ve ilişkisiz değişkenlerin çıkarılmasıyla birlikte modeli açıklayan değişkenler elde edilmektedir. Bu yöntem ile hem kaynak ihtiyacı gereksinimi azalırken hem de daha az değişken kullanılarak bilgi kaybı yaşanmadan daha hızlı ve anlaşılır biçimde kurulan modelin başarımı bilgi kaybı yaşanmadan artmaktadır. Bu amaçla çalışmada ilk olarak veri boyutunun düşürülmesi için değişken seçimi yapılmıştır ve daha sonra kullanılacak yöntemler için gerekli parametre seçimleri yapılarak elde edilen daha düşük boyutlu verilere mutajenisite belirleme sürecini iyileştirebilmek adına çeşitli sınıflama yöntemleri uygulanmıştır. Söz konusu yaklaşım, kullanılarak deneyler ile elde edilen mutajenisite bilgisi içeren molekül setine uygulanmış ve sınıflama başarılarında dikkate değer düzeyde artış sınıflama başarıları görülmüştür.

2. Materyal ve Metot

Bir ilacın mükemmel ADME-Tox (Emilim, Dağılım, Metabolizma, Eliminasyon, Toksisite) özelliklerine sahip olarak piyasaya girmiş olması hayati önem taşımaktadır. Toksisitenin ilaç geliştirmenin tüm aşamalarında başlıca başarısızlık nedeni olmasından dolayı, ilaç toksikolojisi klinik öncesi çalışmalardaki

en önemli araştırma alanlarından biri olmuştur. Bu nedenle, daha önce de belirtildiği gibi ilaç keşif sürecinde kilit rol oynayan molekül toksisitesini in-silico tahmin yöntemleri ile önceden belirlemek uzun ve çok masraflı olan ilaç keşif sürecinde zaman ve maliyet bakımından önemli bir tasarruf sağlamaktadır.

Çalışma kapsamında, literatürde bulunan ve moleküllerden oluşan Bursi ve Benchmark veri setleri kullanılmıştır [15, 16]. Bu veri setleri kullanılarak yapılan çalışmalardan farklı olarak 4337 molekülden oluşan Bursi mutajenisite veri seti ile 6512 molekülden oluşan Benchmark veri seti birleştirilmiş ve tekrarlı gözlemler çıkartılmıştır. 10833 moleküle sahip birleştirilmiş veri seti molekül bilgisine ek olarak her bir moleküle ait aktif olup olmama (aktivite) bilgisini de içermektedir. Moleküllere ilişkin kimyasal tanımlayıcıların (molecular descriptors) hesaplanmasında ise kimyasal hesaplama ve moleküler modelleme aracı olan Molecular Operating Environment (MOE) programı kullanılmıştır [17]. MOE aracılığı ile 2 boyutlu kantitatif yapı-aktivite ilişki (2D QSAR) değişkenleri hesaplanmıştır. Bu değişkenler genel olarak; alt bölümlere ayrılmış yüzey alanı (subdivided surface area), bitişiklik ve uzakllık matrisi (adjacency and distance matrix), farmakofor özellikleri (pharmacophore features), kısmi yükler (partial charges), KierHall bağlantısı ve Kappa şekil endeksleri (Kier&Hall connectivity and Kappa shape indices), atom ve bağ sayıları (atom and bond counts), Hueckel teorisi tanımlayıcıları ve fiziki özellikler (physical properties) olmak üzere çeşitli gruplardan gelmektedir [17]. Sonuç olarak elde edilen veri seti, tahmin edilmek istenen mutajenisite bilgisine sahip olan aktivite değişkeni ile birlikte toplamda 193 değişkenden oluşan büyük bir veri seti haline dönüşmüştür. Veri seti, rastgele bir biçimde %70'i eğitim ve %30'u da sınama seti olacak şekilde ayrılarak aktivite bilgisine göre mutajen ve non-mutajen olarak etiketlendiğinde dağılım Tablo 1'de verildiği gibi özetlenmiştir.

Tablo 1. Veri setinin sınıflara göre eğitim ve sınama setindeki dağılımı.

	Mutajen	Non-mutajen	Toplam
Eğitim	4110	3473	7583
Sınama	1792	148	3250
Toplam	5902	4931	10833

2.1. Sınıflama Algoritmaları

Sınıflama yöntemlerinin ifadelerinde kolaylık açısından yöntemlerin literatürde yaygın kullanılan isim ve kısaltmaları tercih edilmiştir. Buna göre, çalışmada geçen rasgele orman (random forest) RF, aşırı rasgeleleştirilmiş ağaçlar (extremly randomized trees) ExtraTrees, aşırı gradyan arttırma (extreme gradient boosting) XGBoost, hafif gradyan arttırma LightGBM, torbalama (bagging) olarak kullanılmıştır.
Çalışmada sınıflama için ağaç tabanlı yöntemler tercih edilmiştir. Karar ağaçları sınıf bilgisi bulunan veriden tümevarım yöntemi ile çıkarım yaparak öğrenen ve ağaç şekline benzer yapısı olan denetimli bir öğrenme algoritmasıdır. Tahmin edilmek istenen hedef değişkenin ölçüm türüne göre regresyon ve sınıflama problemleri için kullanılabilir.

Rastgele orman (RF), 2001 yılında Leo Breiman tarafından önerilmiş bir karar ağacı algoritmasıdır. Bootstrap örnekleme kullanarak veri içerisinden yeni alt veri setleri oluşturur ve bu her alt veri seti için bir ağaç oluşturulur. Oluşturulan bu ağaçlarda yeni düğüm belirlenirken seçili kritere (gini, entropi) göre en iyi bölünme için seçim yapılarak ve yeni dal açılır. Bu seçim sırasında en çok kazancı sağlayan değişken hesaplandığı için eğitim süresi uzun sürmektedir. Bootstrap kullandığı için varyansı düşük ve aşırı öğrenmeye meyilli olmayan modeller oluşturur [18].

2006 yılında Geurts ve arkadaşları tarafından önerilen ExtraTrees algoritması RF algoritmasına benzer olup ağaç yapısını daha rassal hale getirir. RF algoritmasında düğüm seçimi sırasında uygun bölünme noktası için belirlenen kritere göre kazanç hesaplanırken, ExtraTrees için bu seçim rastgele yapılır. Böylece ağaç yapısındaki çeşitlilik artar ve bu şekilde değişkenliğin düşürülmesi amaçlanır. Rastgele seçim yapması sebebiyle de ara bir hesaplama süreci olan bilgi kazancı veya gini indeksi hesabı olmadığı için RF algoritmasına göre göre daha hızlı çalışan bir algoritmadır [19].

Bagging sınıflayıcısında ise bootstrap örnekleme yöntemi ile veri seti içerisinden rastgele iadeli örneklemler çekilir ve bu şekilde alt veri setleri oluşturulur. Oluşturulan bu alt veri setlerine tekli karar ağaçları uygulanır. Elde edilen sonuçlar içerisinden sınıflama için çoğunluk oylaması regresyon için ise ortalamaya göre sonuç belirlenir [20].

Adaboost, Freund ve Schapire tarafından 1996 yılında önerilmiş bir artırım (boosting) algoritmasıdır. Zayıf öğrenicileri bir arada kullanarak güçlü bir öğrenici elde etmeyi amaçlar. Bu yöntemde başlangıçta her bir gözleme aynı ağırlık değeri atanır ve modelin eğitimine başlanır. Her bir iterasyon sonrasında ilgili gözleme ait sonuç doğru ise atanan ağırlık azaltılır yanlış ise ağırlık arttırılır. Belirlenen ağaç sayısına ulaşıncaya kadar modelin eğitimi bu şekilde devam eder [21].

LightGBM (LGB), gradyan artırım algoritmasının farklı bir uygulamasıdır. 2017 yılından beri Microsoft tarafından açık kaynak olarak geliştirilmektedir. Benzer kütüphanelere göre çok daha hızlı çalışmaktadır. Sürekli değişkenlerin histogramlarını çizip kesikli hale getirerek model performansını arttırır ve eğitim süresini kısaltır [22].

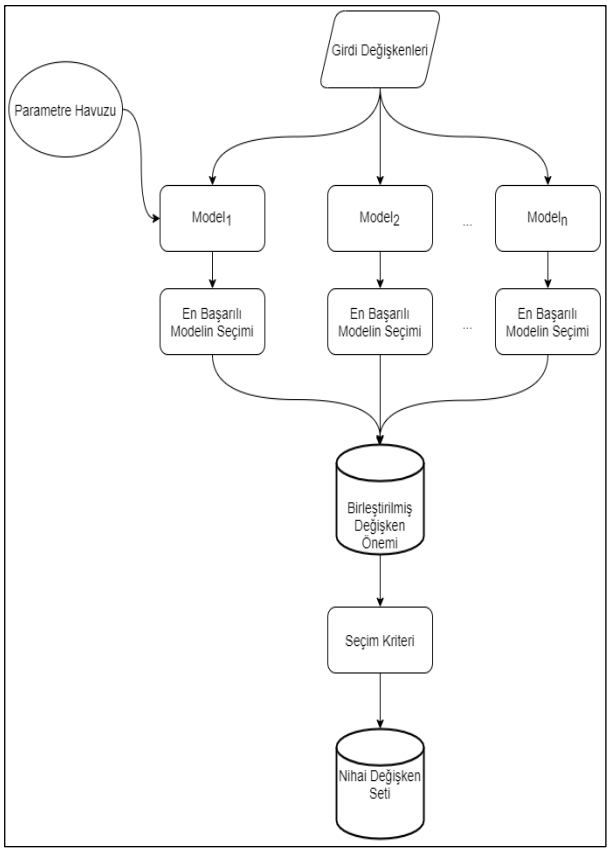
XGBoost (XGB), Chen ve Guestrin tarafından 2014 yılında geliştirilmiş ve devamında dağıtık makine

öğrenmesi topluluğu altında açık kaynak olarak geliştirilmeye devam edilen bir diğer gradyan artırım algoritmasıdır [23].

3. Bulgular

Çalışma kapsamında gerçekleştirilen değişken seçimi, model eğitim ve sınama işlemleri için Python programlama dili kullanılmıştır. Uygulamalarda veri düzenleme işlemleri için pandas, istatistiksel öğrenme yöntem ve değerlendirmeleri için ise scikit-learn, xgboost ve lightgbm kütüphanelerinden yararlanılmıştır.

Uygulamanın ilk aşaması olan değişken seçimi için ağaç tabanlı algoritmalar kullanılarak her bir değişkenin önem düzeyleri belirlenmiştir. Buna göre AdaBoost, ExtraTrees ve RF yöntemleri grid arama (grid search) yöntemi ile çeşitli parametreler ile denenerek en iyi sonuç veren modeller elde edilmiştir. Değişken seçimi Şekil 1'de verilen akıştaki aşamalar sonucu gerçekleştirilmiştir.



Şekil 1. Değişken seçim süreci.

Değişken seçiminin ilk aşaması ExtraTrees, RF ve AdaBoost algoritmaları için ağaç sayısı, öğrenme katsayısı, en küçük ayrım noktası gibi parametrelerin belirlenmesidir. Bu amaç doğrultusunda sırasıyla her bir algoritma için farklı parametre kombinasyonları ile on katlı çapraz doğrulama kullanılarak modeller oluşturulmuştur. Elde edilen modeller arasından en yüksek çapraz doğrulama skoru sağlayan parametreler ile en iyi modeller kurulmuştur. Bu modellerin sonucunda her bir değişken için elde dilen değişken önem düzeylerinin sıralanarak değerlendirilmesi değişken seçiminin takip eden aşamasını oluşturmaktadır. ExtraTrees, RF ve AdaBoost algoritmalarından gelen ve önem düzeylerine göre sıralanan üç ayrı değişken seti tekrar eden değişkenler çıkarılarak birleştirilmiştir. Belirlenen eşik değeri olan ortalama değişken düzeyi üzerinde kalan ve tekrar eden değişken isimlerinin çıkarılması ile değişken seçimi tamamlanmıştır. Buna göre çalışmada kullanılan 193 adet değişkene sahip veri seti uygulanan değişken seçimi sonrasında ağırlıklı olarak kısmi yükler, alt bölümlere ayrılmış yüzey alanı ve bitişiklik ve uzaklık matrisi değişken gruplarından gelen 72 adet değişkenden oluşan nihai veri setine dönüştürülmüştür.

Seçilen değişkenlerin mutajenisiteyi belirlemede ne kadar etkili olduğunun ortaya koyulabilmesi için uygulamalar hem indirgenmemiş 193 adet değişkene sahip veri seti ile hem de 72 adet değişkene sahip indirgenmiş veri seti ile gerçekleştirilmiştir. Tablo 2-4'te verilen algoritmalar için parametre en iyilemesi yapılarak sonuçlar elde edilmiştir. Tüm değişkenlerin dahil edildiği başlangıç veri setine uygulanan algoritmaların her biri için elde edilen doğruluk, duyarlılık ve özgüllük değerlendirme ölçütleri Tablo 2'de verildiği gibidir. Bu tabloya göre, tüm değişkenlerin kullanılması durumunda %88.43 doğruluk yüzdesi ile ExtraTrees ve çok yakın başarı yüzdesi ile RF mutajenisite tahmininde en başarılı algoritmalardır. Sonuç olarak tüm değişkenler kullanılarak %86 ile %88 arasında bir başarım elde edilmiştir.

Tablo 2. Değişken seçimi yapılmadan önceki model başarımları (model parametreleri en iyilenmiştir).

Model	Doğruluk	Duyarlılık	Özgüllük
Bagging	0.8782	0.8608	0.8923
ExtraTrees	0.8843	0.8635	0.8996
Random Forest	0.8834	0.8635	0.8996
LightGBM	0.8603	0.8299	0.8850
XGBoost	0.8812	0.8669	0.8929

Tüm değişkenlerden oluşan veri setine uygulanan yöntemler değişken seçimi sonrasında elde edilen veri setine parametre en iyilemesi yapılarak tekrar uygulanmıştır. Buna göre Tablo 3'te özetlendiği üzere indirgenmiş veri seri ile elde edilen tüm doğruluk oranlarında tüm değişkenleri içeren veri seti doğruluk oranlarına göre artışlar elde edilmiştir. RF algoritması Tablo 2'deki başarı yüzdesi ile karşılaştırıldığında indirgenmiş veri için %1,6 artışla yaklaşık %90 doğruluk başarımına ulaşmıştır. Benzer biçimde Tablo 3 incelendiğinde, mutajenisite tahmininde daha etkili olduğu düşünülen 72 değişken ile elde edilmiş başarı yüzdelerinde tüm değişkenleri kullanarak elde edilen başarı yüzdelerine göre artışlar gözlenmiştir. LightGBM algoritmasının doğruluk oranındaki %3'lük artış da dikkat çekicidir. Her iki veri seti ile elde edilen sonuçları daha net göstermek adına bu farklar Tablo 4'te verilmiştir.

Tablo 4 incelendiğinde değişken seçimi sonrasında tüm modellerde ortalama %1.65'lik başarı artışı görülmektedir. Toksisiteye yol açması nedeniyle oldukça önemli olan mutajenisitenin etkili değişkenler ile birlikte belirlenmesinde in-silico yaklaşım ile elde edilen %1,65'lik artış önemli bir artış olarak değerlendirilebilir.

Tablo 3. Değişken seçimi sonrası model başarımları (Model parametreleri en iyilenmiştir).

Model	Doğruluk	Duyarlılık	Özgüllük
Bagging	0.8904	0.8662	0.9095
ExtraTrees	0.8892	0.8703	0.9045
Random Forest	0.8993	0.8779	0.9168
LightGBM	0.8935	0.8862	0.9157
XGBoost	0.8975	0.8820	0.9101

Tablo 4. Değişken seçimi yapılan veri ve tüm veriye uygulanan modellerin başarım farkları.

Model	Doğruluk (Değişken seçimi)	Doğruluk (Tüm Veri)	Doğruluk (Fark)
Bagging	0.8904	0.8782	0.0122
ExtraTrees	0.8892	0.8843	0.0049
Random Forest	0.8993	0.8834	0.0159
LightGBM	0.8935	0.8603	0.0332
XGBoost	0.8975	0.8812	0.0163

4. Tartışma ve Sonuç

Bu çalışmanın amacı, in-silico yaklaşım kapsamında öğrenme algoritmalarının toksisite testleri arasında ortaya çıkabilecek genetik değişimi yani mutajenisiteyi belirleme sürecinde ne derece etkili olabileceğini ortaya koymaktır. Bu amaç doğrultusunda kullanılan veri seti, mutajenisite aktif olan ve olmayan 10835 molekül için 193 adet değişken, MOE programı ile hesaplatılarak elde edilmiştir. Çalışma değişken seçimi ve mutajenisite tahmini olmak üzere iki genel aşamada gerçekleştirilmiştir. Değişken seçimi aşamasında, veri seti üzerinde ağaç tabanlı algoritmalar olan AdaBoost, ExtraTrees ve RF istatistiksel öğrenme algoritmaları uygulanarak 10 katlı çapraz doğrulama kullanılarak grid arama yaklaşımı ile en iyi parametre seçimi gerçekleştirilmiştir. Elde edilen en iyi parametreler ile kurulan modeller sonucunda ağaç tabanlı modellerden elde edilen değişkenlere ait gini indeksi değerleri kullanılarak mutajenisiteyi tahminlemedeki önem düzeyleri belirlenmiştir. Bu bilgiler ile veri setinin boyutu mutajenisiteyi tahminlemede en etkili 72 değişkene indirgenmiştir. Çalışmada seçilen değişkenler incelendiğinde 72 etkili değişkenin çoğunluğunun (yaklaşık %70) kısmi yükler, alt bölümlere ayrılmış yüzey alanı ve bitişiklik ve uzakllık matrisi değişkenlerini içeren gruplardan geldiği gözlenmektedir. Fiziksel özellikler içeren gruptan reaktif grupların varlığını gösteren özellik ile potansiyel olarak toksik grupların varlığını gösteren özellik mutajenisite tahmininde model tarafından belirlenen en etkili özelliklerin başında gelmektedir. Mutajenisite tahmin aşamasında, seçilen bu tanımlayıcı değişkenlerden oluşan yeni veri setine doğrusal ve doğrusal olmayan 19 farklı istatistiksel öğrenme algoritması uygulanmış, aralarından en iyi sonuç veren beş topluluk öğrenme algoritması seçilmiştir. Seçilen algoritmalar olan Bagging Extra Trees, LightGBM, RF ve XGBoost için yeni değişkenlerden oluşan veri seti kullanılarak tekrar parametre en iyilemesi yapılmıştır. Değişken seçimiyle birlikte yaklaşık %63 oranında boyut indirgemesi yapılmış veri setlerine uygulanmış, değişkenlerin tamamı ile elde edilen sonuçlara göre model başarımlarında %1-3 arası artışlar gözlenmiştir. Sonuç olarak değişken seçimi ve parametre en iyilemesi ile modellerin çoğunda %90'a varan başarım elde edilmiştir.

Bu sonuçlar göstermiştir ki, toksisiteye yol açması nedeniyle özellikle erken evrelerde belirlenmesi çok önemli olan mutajenisite için laboratuvar ortamında (in vitro) ve canlı üzerinde (in vivo) yapılan uzun ve maliyetli çalışmalar öncesinde in-silico yaklaşımlar ile oldukça önemli bulgular elde edilebilmektedir. Günümüz teknolojisi göz önünde bulundurulduğunda bu tür alanlarda in-silico yaklaşımların tercih edilmesi zaman, maliyet ve iş gücündeki azalma nedeni ile çok daha yaygınlaşacaktır.

Çalışmada, değişken seçimi için ağaç tabanlı yöntemler tercih edilmiştir. İleride yapılacak çalışmalarda, değişken seçimi filtre ve sarmal (wrapper) yöntemlere ek olarak L1 (Manhattan) ve L2 (Öklid) düzenlileştirmesi kullanan yöntemler olan Lasso ve Ridge regresyon ile de yapılabilir. Çalışmanın birçok aşamasında kullanılan parametre seçiminde tercih edilen grid arama yönteminde seçilen parametre uzayındaki kombinasyonlar ile arama yapıldığı için bazı sınırlamalar bulunmaktadır. Örneğin doğru öğrenme katsayısı seçilmiş olmasına rağmen ağaç sayısı yeterli olarak verilmez ise model, belirlenen parametre uzayında bulunan en iyiyi bulacak, daha iyi sonuç veren bir kombinasyonu deneyemeyecektir. Bu duruma alternatif olarak ise Bayesçi parametre en iyilemesi ile daha iyi sonuç veren parametreler elde edilebilmesi mümkün olmakla birlikte Bayesçi parametre en iyilemesinde de en iyileme işleminin zaman maliyetinin yüksek olduğu unutulmamalıdır.

Teşekkür

Bu çalışma, Mimar Sinan Güzel Sanatlar Üniversitesi Bilimsel Araştırma Projeleri Komisyonu tarafından desteklenmiştir. (Proje No: 2018-30).

Etik Beyanı

Bu çalışmada, "Yükseköğretim Kurumları Bilimsel Araştırma ve Yayın Etiği Yönergesi" kapsamında uyulması gerekli tüm kurallara uyulduğunu, bahsi geçen yönergenin "Bilimsel Araştırma ve Yayın Etiğine Aykırı Eylemler" başlığı altında belirtilen eylemlerden hiçbirinin gerçekleştirilmediğini taahhüt ederiz.

Kaynakça

[1] Honma, M., Kitazawa, A., Cayley, A., Williams, R. V., Barber, C., Hanser, T., Saiakhov, R., Chakravarti, S., Myatt, G. J., Cross, K. P., Benfenati, E., Raitano, G., Mekenyan, O., Petkov, P., Bossa, C., Benigni, R., Battistelli, C. L., Giuliani, A., Tcheremenskaia, O., … Rathman, J. 2019. Improvement of quantitative structure-activity relationship (QSAR) tools for predicting Ames mutagenicity: Outcomes of the Ames/QSAR International Challenge Project. Mutagenesis, 34(1) 41-48.

[2] Bakhtyari, N. G., Raitano, G., Benfenati, E., Martin, T., Young, D. 2013. Comparison of in silico models for prediction of mutagenicity. Journal of Environmental Science and Health - Part C Env. Carcinogenesis and Ecotoxicology Reviews, 31(1), 45–66.

[3] Hansch, C. 1980. Use of quantitative structure-activity relationships (QSAR) in drug design (review). In Pharmaceutical Chemistry Journal 14(10).

[4] Greene, N., Judson, P. N., Langowski, J. J., Marchant, C. A. 1999. Knowledge-based expert systems for toxicity and metabolism prediction: DEREK, StAR and METEOR. SAR and QSAR in Environmental Research, 10:2-3, 299-314.

[5] Hanser, T., Barber, C., Rosser, E., Vessey, J. D., Webb, S. J., Werner, S. 2014. Self organising hypothesis networks: A new approach for representing and structuring SAR knowledge. Journal of Cheminformatics, 6(21).

[6] Mazzatorta, P., Tran, L. A., Schilter, B., Grigorov, M. 2007. Integration of structure - Activity relationship and artificial intelligence systems to improve in silico prediction of ames test mutagenicity. Journal of Chemical Information and Modeling, 47(1), 34–38.

[7] Zheng, M., Liu, Z., Xue, C., Zhu, W., Chen, K., Luo, X., Jiang, H. 2006. Mutagenic probability estimation of chemical compounds by a novel molecular electrophilicity vector and support vector machine. Bioinformatics, 22(17), 2099–2106.

[8] Liao, Q., Yao, J., & Yuan, S. 2007. Prediction of mutagenic toxicity by combination of Recursive Partitioning and Support Vector Machines. Molecular Diversity, 11, 59–72.

[9] Xu, C., Cheng, F., Chen, L., Du, Z., Li, W., Liu, G., Lee, P. W., Tang, Y. 2012. In silico prediction of chemical ames mutagenicity. Journal of Chemical Information and Modeling, 52(11), 2840–2847.

[10] Moorthy, N. H. N., Kumar, S., Poongavanam, V. 2017. Classification of carcinogenic and mutagenic properties using machine learning method. Computational Toxicology, 3, 33-43.

[11] Zhang, H., Kang, Y. L., Zhu, Y. Y., Zhao, K. X., Liang, J. Y., Ding, L., ... Zhang, J. 2017. Novel naïve Bayes classification models for predicting the chemical Ames mutagenicity. Toxicology in Vitro, 41, 56-63.

[12] Webb, S. J., Hanser, T., Howlin, B., Krause, P., Vessey, J. D. 2014. Feature combination networks for the interpretation of statistical machine learning models: application to Ames mutagenicity. Journal of cheminformatics, 6(1), 1-21.

[13] Seal, A., Passi, A., Jaleel, U. A., Wild, D. J., Open Source Drug Discovery Consortium. 2012. In-silico predictive mutagenicity model generation using supervised learning approaches. Journal of cheminformatics, 4(1), 10.

[14] Ji, X., Tong, W., Liu, Z., Shi, T. 2019. Five-feature Model for Developing the Classifier for Synergistic vs Antagonistic Drug Combinations Built by XGBoost. Frontiers in Genetics, 10, 1-13.

[15] Hansen, K., Mika, S., Schroeter, T., Sutter, A., Ter Laak, A., Steger-Hartmann, T., ... Müller, K. R. 2009. Benchmark data set for in silico prediction of Ames mutagenicity. Journal of chemical information and modeling, 49(9), 2077-2081.

[16] Kazius, J., McGuire, R., Bursi, R. 2005. Derivation and validation of toxicophores for mutagenicity prediction. Journal of medicinal chemistry, *48*(1), 312-320.

[17] MOE, Molecular Operational Environment. Chemical Computing Group Inc., Montreal, Canada.

[18] Breiman, L., 2021. Random forests. Maching Learning, 45(1), 5–32.

[19] Geurts, P., Ernst, D., Wehenkel, L. 2006. Extremely randomized trees. Machine learning, 63(1), 3-42.

[20] Breiman, L. 1996. Bagging predictors. Machine learning, 24(2), 123-140.

[21] Freund, Y., Schapire, R. E. 1996. Experiments with a new boosting algorithm. Machine Learning: Proceedings of the Thirteenth International Conference, July 1996, Italy 148-156.

[22] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... Liu, T. Y. 2017. Lightgbm: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems, Dec 4-9, Long Beach, CA 3146-3154.

[23] Chen, T., Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2016, San Fransisco, California, 785-794.