

Madde Bağımlılığı Riskinin Tahmininde Ridge Classifier Yaklaşımı: Bayeşçi Hiperparametre Araması ve Sınıf Dengesizliği İncelemesi

Çağrı Karakuş¹, Tankut Arca Can¹, Engin Çetintaş¹ and Velihan Doğrul¹

Keywords: Madde Bağımlılığı, Ridge Classifier, Makine Öğrenmesi, Bayesian Optimization, Sınıflandırma

Abstract

Substance addiction is an important public health problem that negatively affects physical and psychological health and causes social and economic problems. This study analyses in detail the performance of the Ridge Classifier algorithm on the UCI Drug Consumption dataset in order to detect addiction risk at an early stage. The value of α , which is the basic hyperparameter of the model, was determined using Bayesian optimisation (BayesSearchCV) approach. In the data preprocessing stage, missing value management, quantification of categorical variables and standardisation of features were applied. Then, the Ridge Classifier was trained for each item and the accuracy, F1 score and class-wise precision-recall values were reported. The results show that the class imbalance problem reduces the model performance in the ‘medium’ and ‘high’ usage classes, while stable performance is obtained in the dominant classes. In particular, it is observed that optimising the α parameter with Bayesian search increases the generalisability of the Ridge Classifier. The study emphasises the importance of using techniques such as data oversampling and class weighting in future research.

Özet

Madde bağımlılığı, fiziksel ve psikolojik sağlığı olumsuz etkileyerek toplumsal ve ekonomik sorunlara yol açan önemli bir halk sağlığı problemidir. Bu çalışma, bağımlılık riskini erken dönemde tespit edebilmek amacıyla UCI Drug Consumption veri seti üzerinde Ridge Classifier algoritmasının performansını detaylı biçimde ele almaktadır. Modelin temel hiperparametresi olan α değeri, Bayeşçi optimizasyon (BayesSearchCV) yaklaşımı kullanılarak belirlenmiştir. Veri ön işleme aşamasında eksik değer yönetimi, kategorik değişkenlerin sayısallaştırılması ve özelliklerin standartlaştırılması uygulanmıştır. Ardından, her bir madde için Ridge Classifier eğitilerek doğruluk, F1 skoru ve sınıf bazında precision-recall değerleri raporlanmıştır. Sonuçlar, sınıf dengesizliği sorununun “orta” ve “yüksek” kullanım sınıflarında model başarısını düşürdüğünü gösterirken, baskın sınıflarda istikrarlı performans elde edilmiştir. Özellikle Bayeşçi arama ile α parametresinin optimize edilmesinin Ridge Classifier’in genellenebilirliğini artırdığı gözlemlenmiştir. Çalışma, veri artırma (oversampling) ve sınıf ağırlığı (class weight) gibi tekniklerin gelecekteki araştırmalarda kullanılmasının önemini vurgulamaktadır.

1. Giriş

Madde bağımlılığı, bireyin hem fiziksel hem de zihinsel sağlığını olumsuz etkileyebilen ve toplumsal açıdan ciddi zararlara yol açabilen bir halk sağlığı sorunudur. Erken dönemde bağımlılık riskini tespit edebilmek, koruyucu sağlık hizmetlerinden rehabilitasyon önlemlerine kadar pek çok alanda stratejik müdahalelerin zamanında hayata geçirilmesi açısından kritiktir. Günümüzde, makine öğrenimi yöntemleri madde bağımlılığı riskinin öngörülmesinde giderek daha yaygın kullanılmaktadır [2, 5].

Kullanılan yöntemler arasında Destek Vektör Makineleri (SVM), Karar Ağaçları, Naive Bayes ve Yapay Sinir Ağları yer almakta olup, Ridge Classifier gibi düzenlestirilmiş (regularized) modeller de

son dönemde dikkat çekmektedir [1, 4]. Bu model, lojistik regresyona benzer bir mantıkla çalışmasına karşın, aşırı uyma (overfitting) riskini L2 penaltısı kullanarak azaltır. Dolayısıyla sınırlı veri boyutlarında veya çoklu sınıf (multiclass) yapılarında istikrarlı performans sergileme potansiyeline sahiptir.

2. Kullanılan Veri Seti

Bu çalışmada, UCI Drug Consumption veri seti [3] kullanılmıştır. Veritabanı 1885 katılımcı için kayıtlar içermektedir. Her katılımcı için 12 özellik bilinmektedir: NEO-FFI-R (nevrotiklik, dışadönüklük, deneyime açıklık, uyumluluk ve vicdanlılık), BIS-11 (dürtüsellik) ve ImpSS (sansasyon arayışı), eğitim düzeyi, yaş, cinsiyet, ikamet edilen ülke ve etnik kökeni içeren kişilik ölçümleri. Tüm girdi özellikleri başlangıçta kategoriktir ve sayısallaştırılmıştır. Nicelleştirmeden sonra tüm girdi özelliklerinin değerleri gerçek değerli olarak kabul edilebilir. Buna ek olarak, katılımcılara 18 yasal ve yasadışı uyuşturucu (alkol, amfetamin, amil nitrit, benzodiazepin, esrar, çikolata, kokain, kafein, crack, ecstasy, eroin, ketamin, yasal uçucular, LSD, metadon, mantar, nikotin ve uçucu madde kullanımı ve aşırı talepte bulunanları belirlemek için tanıtılan bir hayali uyuşturucu (Semeron) kullanımı ile ilgili sorular sorulmuştur. Her bir uyuşturucu için şu cevaplardan birini seçmeleri gerekmektedir: uyuşturucuyu hiç kullanmadım, on yıldan uzun bir süre önce kullandım veya son on yıl, yıl, ay, hafta veya gün içinde kullandım; ancak bu çalışma kapsamında bu seviyeler "0"(düşük ihtimal) "1"(ihtimal bulunduruyor) "2"(yüksek ihtimal) olarak yeniden kodlanmıştır.

Özellikle “yüksek kullanım” seviyesinde bazı maddelerin son derece az gözlem barındırması, sınıf dengesizliği problemini beraberinde getirmektedir. Bu durum, modelin antrenman ve doğrulama aşamalarında kritik öneme sahiptir.

3. Materyal ve Yöntem

3.1. Veri Ön İşleme Adımları

Veri setinde eksik gözlem bulunmamakla birlikte, nadir durumlarda karşılaşılan tutarsız veriler silinmiş ya da bu veriler ortalama veya medyan değerleriyle ikame edilmiştir. Cinsiyet, etnik köken ve eğitim gibi kategorik değişkenler label encoding yöntemi kullanılarak sayısal değerlere dönüştürülmüştür. Değişkenler arasındaki farklı ölçekleri uyumlu hale getirmek için StandardScaler uygulanmış ve böylece tüm özellikler, ortalaması 0 ve standart sapması 1 olacak şekilde standartlaştırılmıştır. Ayrıca, sınıflar kodlanarak her bir madde için CL0 ve CL1 “0” (düşük), CL2–CL4 “1” (orta), CL5–CL6 “2” (yüksek) olarak ayrılmış ve bu sayede üç sınıflı bir problem yapısı oluşturulmuştur.

3.2. Ridge Classifier ve Bayes Araması

Ridge Classifier, lojistik regresyonun log-loss yerine kare kayıp fonksiyonunu (squared loss) kullanan ve L2 düzenleme ekleyen bir sınıflandırma yöntemidir. Bu çalışmada, Ridge Classifier modelinin α hiperparametresi, BayesSearchCV (scikit-optimize kütüphanesi) ile optimize edilmiştir. Hiperparametre arama işlemi, α parametresi için 10^{-5} ile 10^3 aralığında log-uniform bir dağılım kullanılarak gerçekleştirilmiştir. Arama yöntemi olarak, 32 iterasyondan oluşan bir Bayes optimizasyonu uygulanmış ve bu süreçte 5 katlı çapraz doğrulama (cv = 5) yöntemi tercih edilmiştir. Modelin başarımını değerlendirmek için hedef metrik olarak accuracy kullanılmıştır.

Hiperparametre optimizasyonu sonucunda, en iyi α değeri belirlendikten sonra, model bu değeri kullanarak yeniden eğitilmiş ve eğitim seti üzerinde nihai olarak optimize edilmiştir. Ardından, modelin genel performansını değerlendirmek için test seti üzerindeki doğruluk metriği hesaplanmıştır. Bu süreç, Ridge Classifier’ın parametrelerinin doğru bir şekilde ayarlanarak modelin genelleme performansını artırmayı hedeflemiştir.

4. Deneysel Bulgular ve Tartışma

4.1. Genel Performans Değerlendirmesi

Her bir madde için elde edilen en iyi α değeri ve sınıflandırma istatistikleri (doğruluk, precision, recall, f1-score) Tablo 1’te özetlenmiştir. Örneğin, alkol maddesinde $\alpha \approx 19.8457$ seçilmiş ve test setinde doğruluk 0.663 bulunmuştur. Buna karşın “orta kullanım” sınıfında *recall* düşük kalmıştır (%11), bu da veri setinde “orta” kullanımın kısıtlı sayıda gözlem içermesinden kaynaklanmaktadır.

Table 1: Her bir madde için Bayes Araması ile elde edilen en iyi α değeri ve test seti performans metrikleri

Madde	Best α	Accuracy	Precision	Recall	F1-Score
Alcohol	19.85	0.663	0.64	0.66	0.57
Amphet	93.50	0.675	0.61	0.67	0.64
Amyl	50.02	0.807	0.65	0.81	0.72
Benzos	0.0191	0.638	0.56	0.64	0.60
Caff	0.0191	0.869	0.76	0.87	0.81
Cannabis	50.02	0.601	0.62	0.60	0.53
Choc	0.0191	0.813	0.66	0.81	0.73
Coke	440.38	0.657	0.61	0.66	0.62
Crack	0.0191	0.894	0.80	0.89	0.84
Ecstasy	0.0191	0.700	0.67	0.70	0.68
Heroin	0.0191	0.876	0.77	0.88	0.82
Ketamine	24.91	0.813	0.66	0.81	0.73
Legalh	156.17	0.726	0.67	0.73	0.69
LSD	7.4512	0.735	0.71	0.73	0.72
Meth	7.4512	0.758	0.62	0.76	0.67
Mushrooms	0.0191	0.730	0.70	0.73	0.72
Nicotine	191.18	0.527	0.51	0.53	0.47
Semer	0.0191	0.995	0.99	0.99	0.99
VSA	0.0191	0.869	0.76	0.87	0.81

4.2. Sınıf Dengesizliği ve Az Gözlemlili Sınıflar

Pek çok maddede “yüksek kullanım” sınıfının destek (support) sayısı oldukça düşüktür. Örneğin, semeron (Semer) adlı maddede toplam test setinde “2” sınıfı sadece 3 örnekten ibarettir. Model, çoğunlukla “kullanılmıyor (0)” sınıfına tahminde bulunarak 0.995 gibi yüksek bir doğruluk elde etmiş görünmektedir; ancak bu durum gerçek bir sınıflandırma başarısı olarak değerlendirilmemelidir. Bu, modelin dengesiz veri setlerinde çoğunluk sınıfına yönelme eğiliminin örneğidir.

4.3. Confusion Matrix ve Katsayı Analizi

Her madde için elde edilen confusion matrix (Şekil 1) incelendiğinde, modelin “orta kullanım” sınıfındaki hatalarının nispeten yüksek olduğu, “kullanılmıyor (0)” sınıfını doğru tahmin etme oranının ise oldukça yüksek seyrettiği gözlemlenmiştir.

Ayrıca, Ridge Classifier’ın her bir maddeye yönelik *coef_* vektörleri (özellik katsayıları) farklılık göstermektedir. Örneğin, “Impulsive” ve “SS” (duygu arayışı) katsayılarının kokain kullanımında pozitif ve büyük değerler alması, bu özelliklerin kokain kullanım seviyesini tahmin etmede model açısından önemli olduğunu işaret etmektedir.

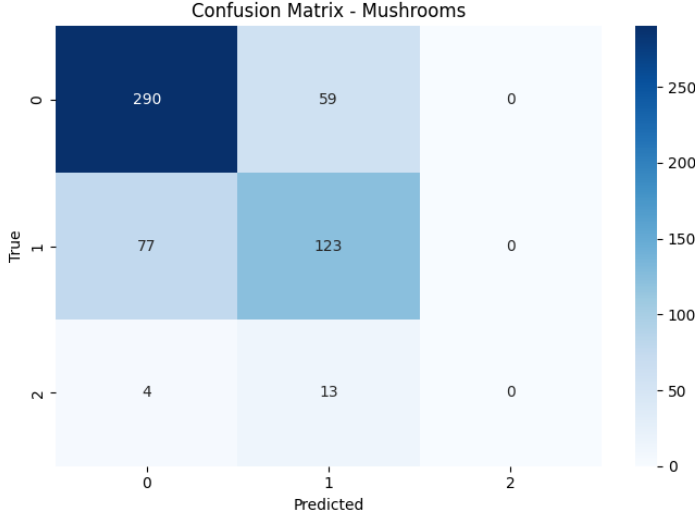


Figure 1: Mantar maddesine ilişkin confusion matrix *confusion matrix* . **Sınıf 0:** Düşük kullanım, **Sınıf 1:** Orta kullanım, **Sınıf 2:** Yüksek kullanım.

5. Tartışma

Ridge Classifier, L2 düzenleme kullanarak aşırı öğrenme riskini azaltmış ve farklı sınıfları makul bir doğrulukla ayırt edebilmiştir. Modelin α parametresinin belirlenmesinde kullanılan Bayesçi optimizasyon yöntemi, hiperparametrenin veri setine uyumlu bir şekilde seçilmesini sağlamıştır. Ancak, bazı durumlarda α değerinin aşırı küçük ($0.0191 \sim 0.0191$) veya büyük ($440 \sim 440$) olması, veri setinin sınıf dağılımı ve özellik uzayının farklı karakteristikleriyle ilişkilendirilebilir. Bu durum, verinin karmaşıklığını ve sınıflar arasındaki ayrımın modelin öğrenme kapasitesini zorlayabileceğini göstermektedir.

Bununla birlikte, veri setindeki sınıf dengesizliği, modelin performansını olumsuz etkileyen önemli bir sorun olarak ortaya çıkmaktadır. Özellikle, az sayıda örneğe sahip olan “yüksek kullanım” sınıflarında modelin precision ve recall değerlerinin sıklıkla 0.00 olması dikkat çekicidir. Bu durum, modelin bu sınıfları neredeyse hiç tahmin edemediğini ve sınıf dengesizliğinin etkilerini açıkça yansıttığını göstermektedir.

Bu sorunu hafifletmek ve modelin daha dengeli bir performans göstermesini sağlamak için çeşitli stratejiler önerilmektedir. İlk olarak, modelin eğitimi sırasında sınıf ağırlıklarının `{class_weight="balanced"}` olarak otomatik şekilde belirlenmesi veya manuel olarak uygun ağırlıkların atanması, nadir sınıflara daha fazla önem verilmesini sağlayabilir. Bu yöntem, sınıf dengesizliğini hafifletmek için etkili bir yol sunar.

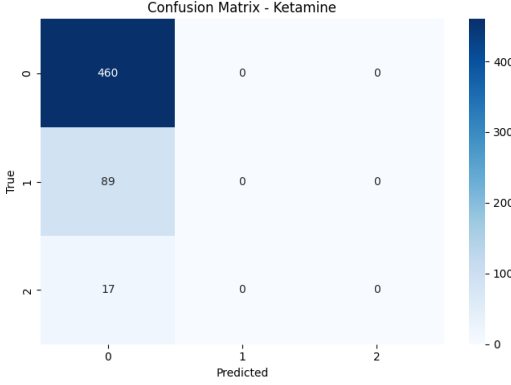
Bir diğer yaklaşım, veri çoğaltma yöntemlerinden yararlanmaktır. Özellikle SMOTE (Synthetic Minority Over-sampling Technique) gibi teknikler kullanılarak “yüksek kullanım” sınıflarına ait sentetik örnekler oluşturulabilir. Bu sayede, nadir sınıfların temsil gücü artırılarak modelin bu sınıfları daha iyi öğrenmesi sağlanabilir.

Son olarak, sınıfların yeniden etiketlenmesi gibi hedef dönüşümü yöntemleri de dikkate alınabilir. Örneğin, orta ve yüksek kullanım sınıflarını birleştirerek ikili sınıflandırma problemi haline getirmek, sınıf dengesizliğinin etkilerini azaltabilir. Bu tür bir yaklaşım, modelin daha az karmaşık bir problem üzerinde çalışmasını sağlayarak tahmin doğruluğunu artırabilir.

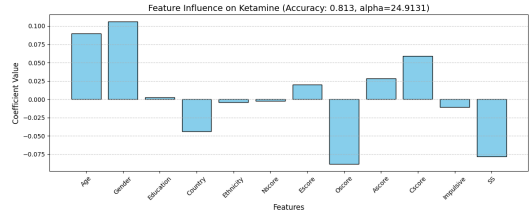
Tüm bu yöntemler, modelin performansını iyileştirmeye ve veri setindeki sınıf dengesizliğinin olumsuz etkilerini azaltmaya yönelik etkili çözümler sunmaktadır.

6. Ketamin (Ketamine) Kullanımı Analizi

6.1. Sınıf Dağılımı ve Performans



(a) Ketamin için confusion matrix.



(b) Ketamin için özelliklerin sonuca etkisi.

Figure 2: Ketamin kullanımı için sınıf dağılımı ve performans metrikleri.

Şekil 2’te gösterildiği gibi, ketamin kullanımı için "0" (kullanılmıyor) sınıfı baskın bir destek sayısına sahiptir. Veri setindeki bu belirgin dengesizlik, modelin diğer sınıflar üzerinde düşük performans göstermesine yol açmıştır. Model, "0" sınıfında yüksek precision ve recall değerleri elde etmiş, ancak "1" (orta kullanım) ve "2" (yüksek kullanım) sınıflarında bu metrikler sıfıra yakın olmuştur.

6.2. Özelliklerin Kullanım Üzerindeki Etkisi

Ketamin kullanımını etkileyen özelliklerin analizi, bireysel ve psikolojik faktörlerin madde kullanımı üzerindeki etkisini anlamak açısından önemli bilgiler sunmaktadır. Yaş değişkeni, ketamin kullanımının tahmin edilmesinde pozitif bir etki göstermektedir; bu durum, yaş ilerledikçe ketamin kullanım olasılığının arttığına işaret etmektedir. Cinsiyet değişkeni ise yüksek pozitif bir etkiye sahiptir ve belirli bir cinsiyet grubunun ketamin kullanımında diğerine göre daha yüksek bir eğilime sahip olabileceğini ortaya koymaktadır. Bunun yanı sıra, deneyime açıklık (Oscore) değişkeni, ketamin kullanımı üzerinde negatif bir etki göstermektedir; düşük deneyime açıklık değerlerinin ketamin kullanımının daha olası olduğu durumlarla ilişkili olduğu gözlemlenmiştir.

Sorumluluk (Cscore) değişkeni, ketamin kullanım tahmininde pozitif bir etkisiyle öne çıkmaktadır; yüksek sorumluluk seviyelerinin ketamin kullanımını artırabileceği görülmektedir. Ayrıca, duygu arayışı (SS) değişkeni de ketamin kullanımı üzerinde negatif bir etki göstermektedir; düşük duygu arayışı düzeylerinin ketamin kullanım olasılığını artırabileceği ortaya konulmuştur. Bu sonuçlar, bireylerin yaş, kişisel özellikler ve psikolojik faktörler gibi unsurların madde kullanımı üzerindeki belirleyici etkisini ortaya koymakta ve tahmin modellerinin geliştirilmesinde kritik içgörüler sunmaktadır.

6.3. Confusion Matrix

Şekil 4’te görüldüğü üzere, model "0" sınıfında yüksek bir doğruluk oranı yakalamış ancak "1" ve "2" sınıflarında yanlış sınıflandırma oranı oldukça yüksektir. Ketamin kullanımını tahmin eden modelin doğruluk oranı

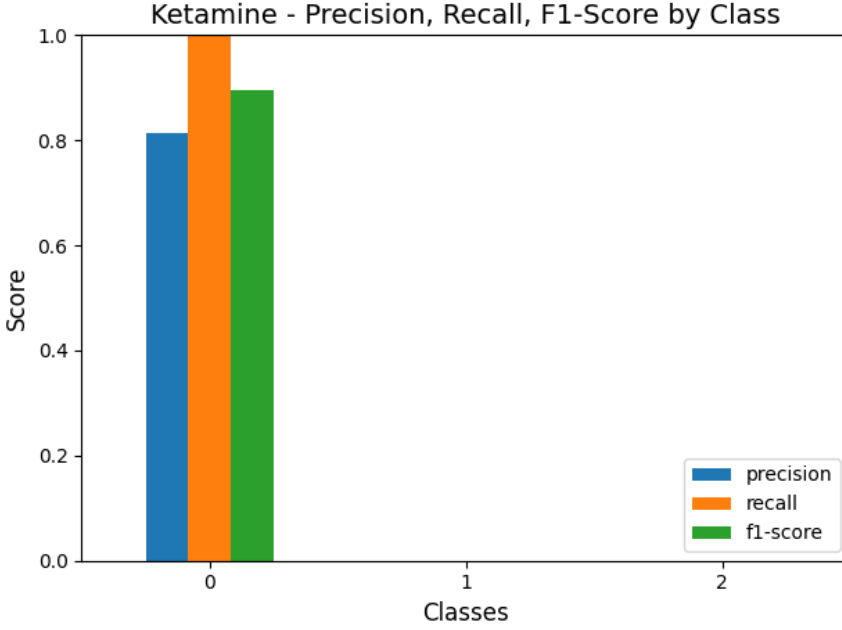


Figure 3: Ketamin için precision - recall - f1score (doğruluk: 0.81).

7. Sihirli Mantar (Mushrooms) Kullanımı Analizi

Bu bölümde, sihirli mantar (mushrooms) kullanımına ilişkin modelleme sonuçları ve analizler yer almaktadır. Modelin doğruluk oranı, sınıf bazındaki performansı ve özellik etkileri detaylandırılmıştır.

7.1. Sınıf Dağılımı ve Performans

Şekil 5'te görüldüğü gibi, "0" (kullanılmıyor) sınıfı veri setinde baskınken, "2" (yüksek kullanım) sınıfı oldukça sınırlı bir örnek içermektedir. Bu dengesizlik, modelin doğruluk oranını ve sınıf bazındaki tahmin performansını etkilemiştir. Özellikle "0" ve "1" sınıflarında precision ve recall değerleri kabul edilebilir seviyedeysen, "2" sınıfı için model hiçbir doğru tahmin yapamamıştır.

7.2. Özelliklerin Kullanım Üzerindeki Etkisi

Modelde etkili olan özellikler, Şekil 6'te sunulmaktadır. Yaş (*Age*) ve ülke (*Country*) gibi değişkenler model tahminlerinde güçlü pozitif etkiler sergilemiştir. Bunun yanı sıra, deneyime açıklık (*Oscore*) negatif bir katkı sağlamış; düşük değerlere sahip bireylerin kullanım tahminlerinde belirleyici olduğu gözlemlenmiştir. Ayrıca, duygu arayışı (*SS*) da negatif etkisiyle dikkat çekmiştir.

7.3. Confusion Matrix Değerlendirmesi

Şekil 7'te verilen confusion matrix, modelin "0" sınıfında doğru tahminler yaptığını, ancak "1" ve özellikle "2" sınıflarında önemli hatalar içerdiğini göstermektedir. "2" sınıfı için hiçbir doğru tahmin yapılmamış, bu durum sınıf dengesizliğinin bir sonucu olarak değerlendirilmiştir. Modelin genel doğruluk oranı

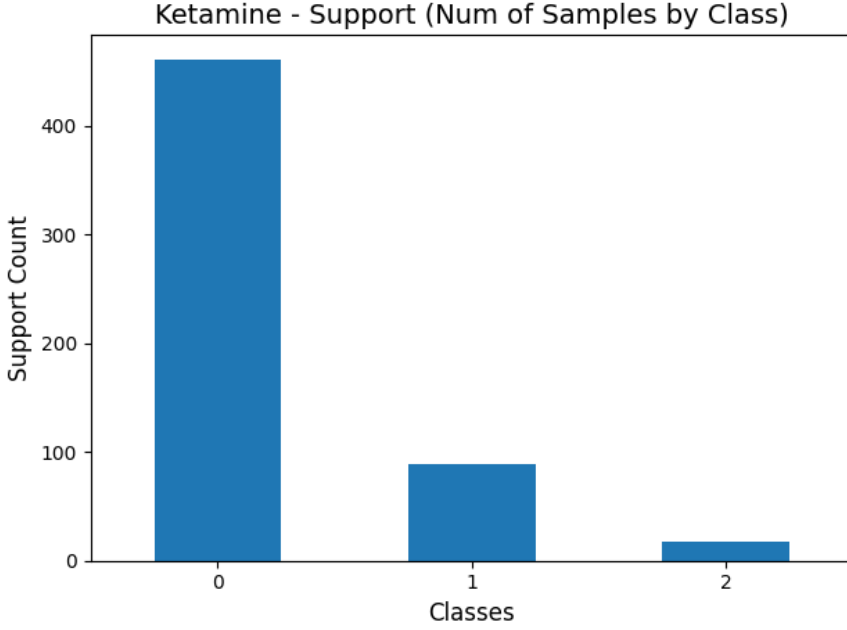
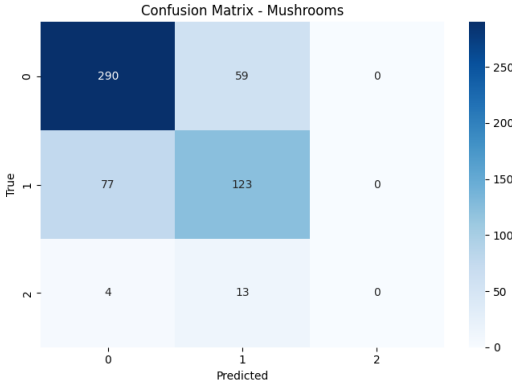
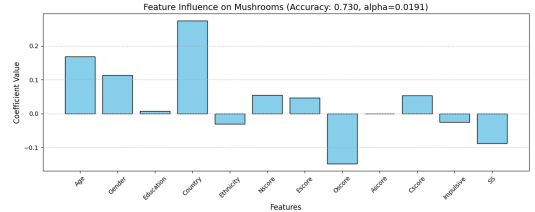


Figure 4: Ketamin kullanımı için sınıflara göre örnekler.



(a) Sihirli Mantar için confusion matrix.



(b) Sihirli Mantar için özelliklerin etkisi.

Figure 5: Sihirli mantar kullanımı için sınıf dağılımı ve performans metrikleri.

8. Benzodiazepin (Benzos) Kullanımı Analizi

Benzodiazepin (Benzos) kullanımına ilişkin analiz, sınıf dağılımı, model performansı ve tahminlerde etkili olan özelliklerin değerlendirilmesini kapsamaktadır. Aşağıdaki bölümlerde bu analizler detaylandırılmıştır.

8.1. Sınıf Dağılımı ve Performans

Benzodiazepin kullanımına dair sınıf dağılımı ve modelin performansını değerlendiren sonuçlar Şekil 8'te sunulmuştur.

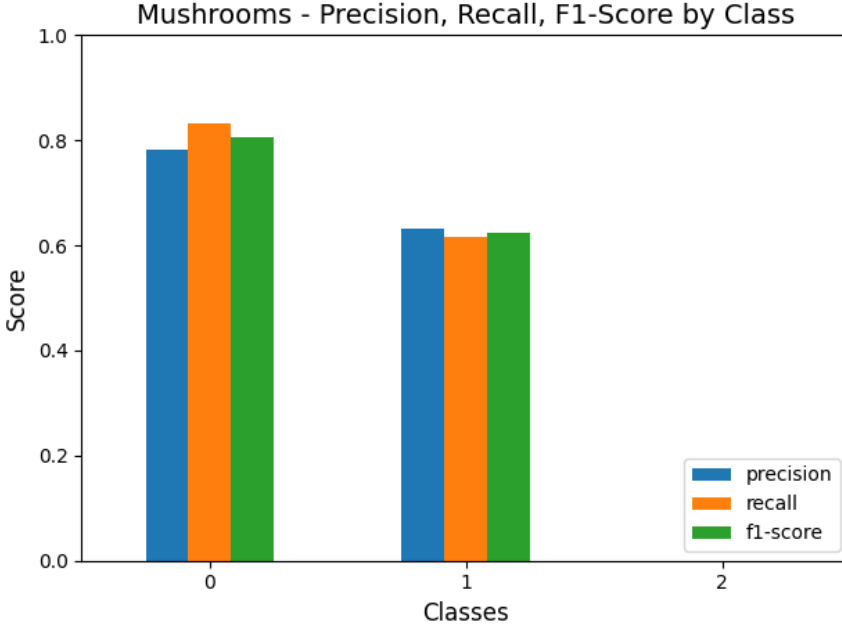


Figure 6: Sihirli mantar için precision-recall-f1score.

Şekil 8a, sınıfların destek (örnek sayısı) dağılımını göstermektedir. "0" (kullanılmıyor) sınıfı, en fazla destek alan sınıf olarak öne çıkarken, "1" (orta kullanım) ve "2" (yüksek kullanım) sınıfları daha az sayıda örnek içermektedir. Bu dengesizlik, modelin tahmin performansını etkilemiştir. Şekil 8b’te modelin performansı değerlendirildiğinde, "0" sınıfında precision (%68) ve recall (%97) oranlarının yüksek olduğu, ancak "1" ve "2" sınıflarında performansın daha düşük olduğu gözlemlenmiştir.

8.2. Özelliklerin Kullanım Üzerindeki Etkisi

Tahminlerde etkili olan özellikler ve modelin doğruluğunu etkileyen faktörler Şekil 9’te sunulmaktadır.

Şekil 9’te gösterildiği gibi, benzodiazepin kullanımı tahminlerinde çeşitli demografik ve davranışsal özellikler önemli bir rol oynamaktadır. Yaş (Age) değişkeni, pozitif bir etkiye sahiptir ve benzodiazepin kullanım tahminlerinde belirleyici bir faktör olarak öne çıkmaktadır. Benzer şekilde, Deneyime Açıklık (Oscore) özelliği de pozitif bir etki göstermekte olup, daha açık deneyime sahip bireylerin kullanım olasılığını artırdığı gözlemlenmiştir.

Diğer taraftan, Sorumluluk (Cscore) özelliği negatif bir etki sergilemektedir; bu durum, daha düşük sorumluluk seviyelerine sahip bireylerde benzodiazepin kullanım olasılığının daha yüksek olabileceğini göstermektedir. Duygu Arayışı (SS) değişkeni de negatif bir etkiye sahip olup, düşük duygu arayışının kullanım tahmini üzerinde etkili olduğunu ortaya koymaktadır. Ayrıca, Ülke (Country) değişkeni, bazı ülkelerde bireylerin kullanım olasılığının daha düşük olduğunu işaret eden negatif bir etkide bulunmuştur.

Modelin doğruluk oranı %63.8 olarak hesaplanmıştır. Benzodiazepin kullanımında "0" sınıfında yüksek doğruluk sergileyen model, diğer sınıflarda veri dengesizliği nedeniyle performans kaybı yaşamaktadır.

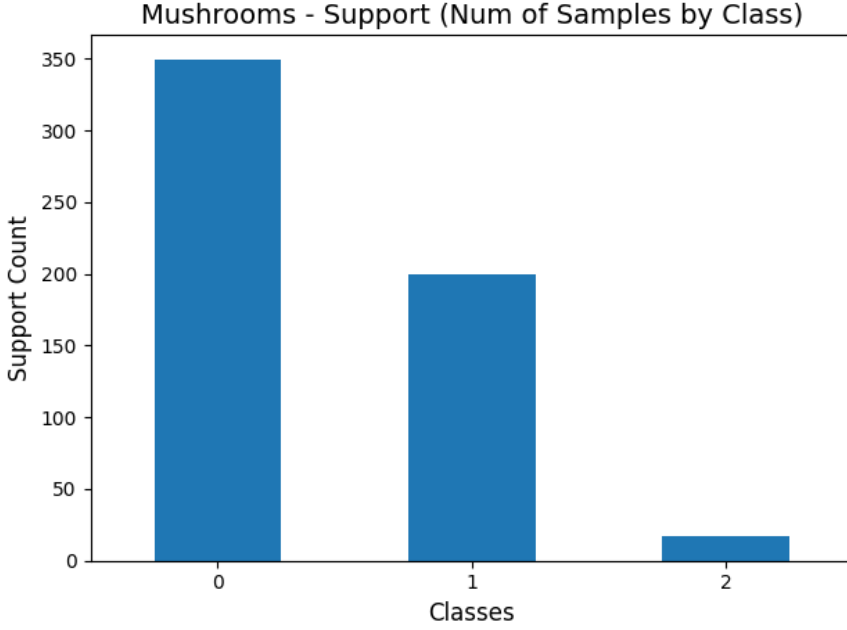
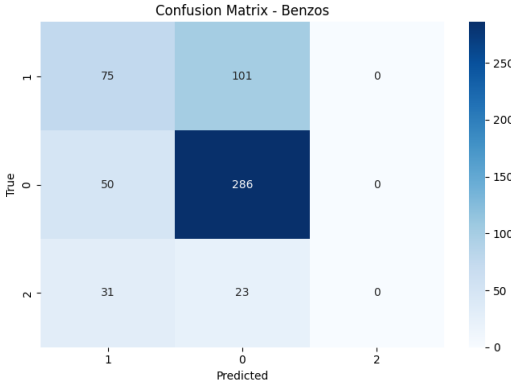
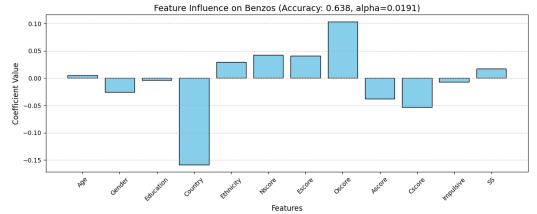


Figure 7: Sihirli mantar kullanımı için sınıflara göre örnekler.



(a) Benzodiazepin için confusion matrix.



(b) Benzodiazepin için özelliklerin etkisi.

Figure 8: Benzodiazepin kullanımı için sınıf dağılımı ve performans metrikleri.

9. Ortak Değerlendirme ve İyileştirme Önerileri

Ele alınan üç farklı madde (ketamin, sihirli mantar ve benzodiazepin) için geliştirilen modellerde ortak bir sorun olarak sınıf dengesizliği ve bunun sonucu olarak yüksek kullanım sınıflarında düşük performans dikkat çekmektedir. Bu genel tablo, aşağıdaki yaklaşımlarla iyileştirilebilir:

Veri dengesizliğini azaltmak amacıyla, düşük örnek sayısına sahip sınıflar için oversampling veya undersampling gibi yöntemler uygulanabilir. Bunun yanı sıra, modelin eğitim sürecinde baskın olmayan sınıflara daha yüksek ağırlık tanımlayan class weight tekniği kullanılarak, sınıflar arasındaki dengesizlik azaltılabilir. Performansın daha da artırılması için, derin öğrenme tabanlı modeller veya ensemble (ensemble) yöntemler gibi gelişmiş modelleme yaklaşımları değerlendirilebilir. Ayrıca, grid search, random search veya bayes optimizasyonu gibi hiperparametre arama teknikleri ile modelin

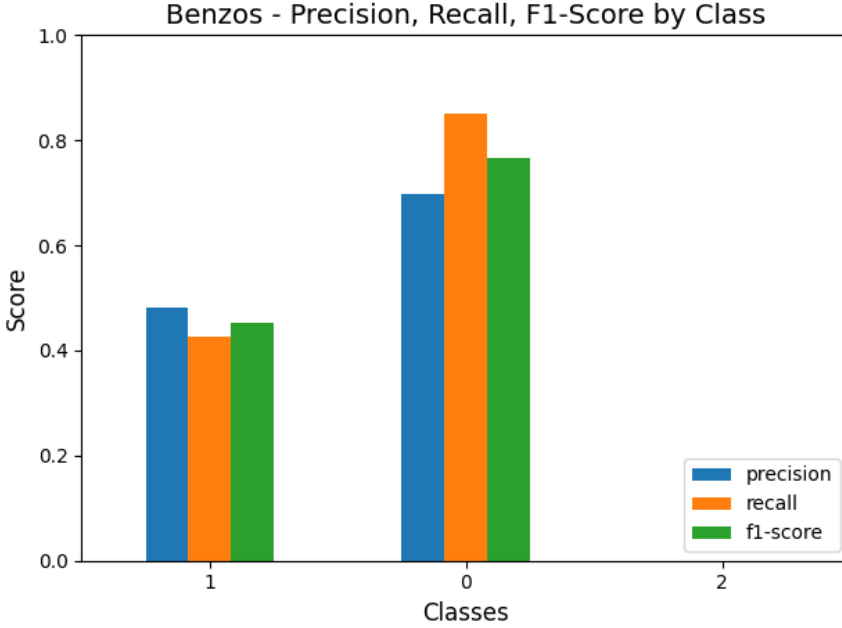


Figure 9: Benzodiazepin için precision-recall-f1score.

en uygun ayarları bulunarak sonuçların iyileştirilmesi sağlanabilir. Son olarak, daha geniş ve dengeli veri setlerinin toplanması veya mevcut verilerdeki sınıfların dengesinin yeniden düzenlenmesi, modelin genellenebilirliğini ve yüksek kullanım sınıflarındaki doğruluğu artırmaya yardımcı olacaktır. Bu öneriler, incelenen tüm maddeler özelinde model performansını ve dengesiz sınıflardaki doğruluğu yükseltmeyi amaçlamaktadır.

10. Sonuç ve Öneriler

Bu çalışma, UCI Drug Consumption veri seti üzerinde Ridge Classifier modelinin performansını, Bayes araması kullanarak incelemiş ve elde edilen sonuçları sunmuştur. Analizler, modelin baskın sınıflarda makul doğruluk ve hassasiyet (precision) değerleri elde ettiğini, ancak “orta” ve “yüksek” kullanım gruplarında veri yetersizliği nedeniyle düşük performans sergilediğini göstermiştir. Bu durum, sınıf dengesizliğinin modelin genelleme yeteneği üzerindeki olumsuz etkilerini açıkça ortaya koymaktadır.

Gelecekteki çalışmalarda, bu sınırlamaların üstesinden gelmek için bazı öneriler değerlendirilebilir. İlk olarak, sınıf dengesizliği problemini ele almak için SMOTE, ADASYN gibi oversampling yöntemleri veya undersampling teknikleri kullanılabilir. Bu yaklaşımlar, özellikle düşük örneklem sayısına sahip sınıfların daha iyi temsil edilmesini sağlayabilir. Ek olarak, sosyal medya verileri, biyometrik ölçümler gibi farklı veri kaynaklarının entegrasyonu, sınıf ayrımını güçlendirebilir ve modelin performansını artırabilir.

Daha gelişmiş modelleme yaklaşımları da dikkate alınabilir. Örneğin, Ridge Classifier yerine L1 (Lasso) düzenleme, ElasticNet veya hiperparametresi farklı olan daha karmaşık modellerin test edilmesi, daha iyi sonuçlar elde edilmesini sağlayabilir. Ayrıca, model açıklanabilirliği üzerinde çalışmak da önemli bir katkı sunabilir. SHAP (SHapley Additive exPlanations) gibi yöntemler kullanılarak, hangi özelliklerin tahmini ne ölçüde etkilediği detaylı bir şekilde incelenebilir.

Sonuç olarak, Ridge Classifier gibi düzenlenmiş modellerin, doğru hiperparametre ayarlarıyla (örneğin, α parametresi için Bayesçi optimizasyon) güçlü performans sergileyebileceği görülmüştür.

Bununla birlikte, dengesiz veri setlerinde ek yöntemlerin devreye alınmasının gerekli olduğu anlaşılmaktadır. Model, özelliklerin farklı kullanım seviyelerindeki etkilerini analiz ederek, bağımlılık riskini tahmin etmede **Escore** ve **SS** gibi davranışsal özelliklerin kritik bir rol oynadığını göstermiştir. Bu bulgular, gelecekteki model geliştirme çalışmalarına ışık tutmaktadır.

Acknowledgments. Bu araştırmanın yürütülmesinde destek veren tüm uzmanlara teşekkür ederiz.

Funding Statement. Bu çalışma, X projesi kapsamında desteklenmiştir. Destekleyen kuruluş: TÜBİTAK (Proje No: 123456).

Competing Interests. Yazarlar, bu araştırma ile ilgili herhangi bir çıkar çatışması beyan etmemektedir.

Data Availability Statement. UCI Drug Consumption veri seti, <https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29> adresinden temin edilebilir.

Ethical Standards. Bu çalışma, katılımcı bilgilerinin anonim olduğu ve ek etik beyan gerektirmeyen açık bir veri seti kullanmıştır.

Author Contributions. Konsept ve tasarım: Ç.K. ve T.A.

Veri analizi: E.Ç.

Metin yazımı: V.D., Ç.K.

Bütün yazarlar, makalenin son halini okumuş ve onaylamıştır.

References

- [1] Donald E. Knuth, *Literate programming*, The Computer Journal **27** (1984), no. 2, 97–111.
- [2] ———, *The T_EX book*, Addison-Wesley Professional, 1986.
- [3] Leslie Lamport, *L^AT_EX: a document preparation system*, 2 ed., Addison Wesley, Massachusetts, 1994.
- [4] Michael Lesk and Brian Kernighan, *Computer typesetting of technical journals on UNIX*, Proceedings of American Federation of Information Processing Societies: 1977 National Computer Conference (Dallas, Texas), 1977, pp. 879–888.
- [5] Frank Mittelbach, Michel Gossens, Johannes Braams, David Carlisle, and Chris Rowley, *The L^AT_EX companion*, 2 ed., Addison-Wesley Professional, 2004.