Yalova Üniversitesi
Mühendislik Fakültesi
Bilgisayar Mühendisliği Bölümü

# Makine Öğrenmesi Yöntemleri İle AIDS Virüsü Hastalık Tahmini

Medine İleyda ERDOĞAN

*Yalova Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, 77100, YALOVA*

**Makale Bilgisi**

*Başvuru:08 /05/2024*

**Anahtar Kelimeler**

*AIDS Hastalık Tahmini*
*Makine Öğrenmesi*
*Öznitelik Seçimi*
*Regresyon Analizi*
*Sınıflandırma Algoritmaları*
*Veri Madenciliği*
*Karar Destek Sistemleri*

## Öz

Son yıllarda, AIDS hastalığının doğru bir şekilde tahmin edilmesi, toplumların sağlık politikalarının ve kaynakların doğru bir şekilde dağıtılmasında büyük önem taşımaktadır. Bu çalışma, çeşitli makine öğrenmesi algoritmalarını kullanarak AIDS hastalığını tahmin etmeyi ve analiz etmeyi amaçlamaktadır. Çalışmamız, ülkelerin sosyo-ekonomik göstergelerini, sağlık hizmetlerine erişim verilerini ve sosyal faktörleri içeren geniş bir veri setini kullanarak AIDS hastalığını tahmin etmek için Naive Bayes, k-Nearest Neighbors, Logistic Regression, Support Vector Machines, Random Forest ve XGBoost gibi yöntemleri kapsamlı bir şekilde değerlendirmiştir. Her bir algoritma, çeşitli değerlendirme metrikleri kullanılarak değerlendirilmiş ve model performansları karşılaştırılmıştır. Bu çalışmanın sonuçları, toplum sağlığı politikalarının geliştirilmesine ve kaynakların etkili bir şekilde dağıtılmasına katkıda bulunacaktır.

## Predicting AIDS Virus Disease Approval with Machine Learning Algorithms

**Keywords**

*AIDS Virus Infection Prediction*

*Machine Learning*
*Feature Selection*
*Regression Analysis*
*Classification Algorithms*
*Data Mining*
*Decision Support Systems*

### Abstract

In recent years, accurate prediction of AIDS disease has been of great importance in the proper distribution of public health policies and resources. This study aims to predict and analyze AIDS disease using various machine learning algorithms. Our study comprehensively evaluates methods such as Naive Bayes, k-Nearest Neighbors, Logistic Regression, Support Vector Machines, Random Forest, and XGBoost to predict AIDS disease using a broad dataset including socio-economic indicators of countries, access to healthcare data, and social factors. Each algorithm is evaluated using various evaluation metrics, and model performances are compared. The results of this study will contribute to the development of public health policies and the effective distribution of resources.

# GİRİŞ (INTRODUCTION)

Son yıllarda, HIV/AIDS salgını, küresel sağlık için ciddi bir tehdit oluşturmaya devam etmektedir. HIV/AIDS'in tespiti ve yayılmasının kontrol altına alınması, dünya genelinde halk sağlığı otoritelerinin ve araştırmacıların öncelikli hedeflerinden biri haline gelmiştir. Erken teşhis ve etkili tedavi stratejileri geliştirmek, hastalığın yayılmasını önlemek ve hasta yaşam kalitesini artırmak için kritik öneme sahiptir. Bu bağlamda, ileri veri analitiği ve makine öğrenme tekniklerinin kullanımı, HIV/AIDS vakalarının daha hızlı ve doğru bir şekilde tespit edilmesine yönelik önemli fırsatlar sunmaktadır.

Makine öğrenme modelleri, büyük veri kümelerindeki karmaşık ilişkileri tanımlayabilme yetenekleri sayesinde, sağlık verilerinden anlamlı öngörüler çıkarabilmektedir. Bu çalışma, HIV/AIDS virüsünün tespitinde makine öğrenme algoritmalarının etkinliğini araştırmayı amaçlamaktadır. Çeşitli sosyo-ekonomik göstergeler, sağlık hizmetlerine erişim verileri ve diğer ilgili faktörler kullanılarak, farklı demografik gruplar için HIV/AIDS vakalarının tahmin edilmesi ve analiz edilmesi hedeflenmektedir.

Araştırmada Naive Bayes, k-En Yakın Komşular, Lojistik Regresyon, Destek Vektör Makineleri, Random Forest ve XGBoost gibi çeşitli makine öğrenme algoritmaları değerlendirilecektir. Bu algoritmaların performansları, doğruluk, hassasiyet, duyarlılık ve F1 skoru gibi çeşitli değerlendirme metrikleri kullanılarak titizlikle incelenecektir.

Her bir algoritmanın performansının çeşitli değerlendirme metrikleri kullanılarak dikkatlice incelenmesiyle, bu çalışma HIV virüs yayılımını etkileyen karmaşık dinamikler hakkında değerli içgörüler sunmayı amaçlamaktadır. Son olarak, bu araştırma HIV/AIDS vakalarının tespitine yönelik makine öğrenme modellerinin potansiyelini ve bu modellerin halk sağlığı politikalarına nasıl entegre edilebileceğini araştırmaktadır. Elde edilen bulguların, HIV/AIDS'in yayılmasını azaltmaya ve genel halk sağlığını iyileştirmeye yönelik hedeflenmiş politikaların ve kaynak dağıtım stratejilerinin geliştirilmesine katkı sağlaması beklenmektedir.

# 1. YÖNTEMLER (METHODS)

Bu çalışma, HIV virüsü tespitini tahmin etmek için farklı makine öğrenmesi algoritmalarının etkinliğini değerlendirmek amacıyla tasarlanmıştır. Araştırmanın temeli, gerçek dünya verileri üzerinde bu algoritmaların performansını ölçmek ve karşılaştırmaktır. Algoritmaların seçimi, literatürde yaygın olarak kabul gören ve HIV tespitini değerlendirmek için uygun olan modellere dayanmaktadır. Bu bölümde, kullanılan veri seti, uygulanan algoritmalar ve değerlendirme metrikleri detaylı bir şekilde açıklanmaktadır.

## 1.1. Veri Seti (Data Set)

Bu çalışmada kullanılan veri seti, HIV virüsü enfeksiyon oranlarına ilişkin istatistikler ve ilgili sosyo-ekonomik göstergeleri içeren kapsamlı bir veri kaynağıdır. Veri seti, uluslararası sağlık kuruluşları, ulusal istatistik ofisleri ve akademik araştırma kurumları tarafından yayımlanan güvenilir kaynaklardan toplanmıştır. Ayrıca, veri seti içerisindeki değişkenler arasında yaş, cinsiyet, kan transfüzyonu durumu, uyuşturucu kullanımı, cinsel ilişki türü, gebelik durumu ve HIV enfeksiyonu riskini etkileyen diğer önemli faktörler yer almaktadır. Bu veri seti, HIV virüsü enfeksiyonlarının tahmin edilmesi ve analiz edilmesi için sağlam bir temel sunmaktadır.

***Tablo 1. AIDS veri seti öznitelikleri ve açıklamaları:***

| Öznitelik | Açıklama |
|---|---|
| Tedavi göstergesi (trt) | Hastanın tedavi göstergesi. |
| Yaş (age) | Hastanın yaşı. |
| Kilo (wtkg) | Hastanın kilosu. |
| Hemofili (Hemo) | Hemofili hastası olup olmadığı. |
| Homoseksüel Aktivite (Homo) | Hastanın homoseksüel birliktelik yaşayıp yaşamadığı. |
| Uyuşturucu Madde (Drugs) | Hastanın hikayesinde uyuşturucu kullanımı. |
| Karnofsky Puanı (karnof) | Hastanın sağlık durumunu 0-100 arasında hesaplayan değer. |
| Oprior | HIV pozitif bireylerin son 175 gün içerisinde Non-ZDV antiretroviral tedavi alıp almadığı. |
| Z30 | Hastanın 175. günden önceki 30 gün içinde Zidovudine (ZDV) tedavisi almadığını gösterir. |
| Preanti | 175. günden önce kaç gün boyunca antiretroviral tedavi aldığını gösterir. |
| Irk (Race) | Bireylerin ırksal demografik özelliklerini belirtmek için kullanılır. |
| Cinsiyet (Gender) | Hastanın cinsiyetini belirtir. |
| Antiretroviral Geçmiş (str2) | Hastanın antiretroviral tedavi geçmişlerini belirtmek için kullanılır. |
| Antiretroviral Tedavi Geçmişi (strat) | Hastanın antiretroviral tedavi alıp almadığını, aldıysa da kaç hafta aldığını belirtir. |

| | |
|---|---|
| Symptomatic Indicator (symptom) | Hastada semptomların görülüp görülmediği. |
| Treatment Indicator (treat) | Hastanın Zidovudine (ZDV) dışındaki diğer antiretroviral tedavileri alıp almadığını gösterir. |
| Offtrt | Hastanın 96 haftadan önce tedaviyi kesip kesmediği bilgisini verir. |
| CD40 | Hastanın başlangıçtaki CD4 hücre sayısını ifade eder. |
| CD420 | Hastanın tedavi başladıktan 20 hafta sonraki CD4 hücre sayısını ifade eder. |
| CD80 | Hastanın tedavi başlangıcındaki CD8 hücre sayısını ifade eder. |
| CD820 | Hastanın tedavi başladıktan 20 hafta sonraki CD8 hücre sayısını ifade eder. |

Bu öznitelikler, HIV virüsü taşıma riskini değerlendirmede kullanılan çeşitli faktörleri yansıtmaktadır. Çalışmamızda, bu özniteliklerin intihar riskini tahmin etmedeki etkinliğini belirlemek için çeşitli makine öğrenmesi teknikleri kullanılmıştır. Bu bilgiler, toplumsal refahı artırmaya yönelik politika ve müdahalelerin geliştirilmesine yardımcı olabilir.

## 1.2. Öznitelik Seçme (Feature Selection)

HIV virüsünün bireyler üzerindeki etkilerini tahmin etmek için veri setimizdeki öznitelikler arasında önemli ilişkileri belirlemek ve model performansını artırmak amacıyla öznitelik seçimi yapacağız. Bu seçim, her bir özniteliğin bireysel tahmini yeteneği ile HIV enfeksiyonu üzerindeki etkisinin ölçülmesiyle gerçekleşecek. Ayrıca, öznitelikler arasındaki korelasyonun değerlendirilmesiyle, birbirleriyle yüksek korelasyonlu olan ancak hedef değişkenle düşük korelasyona sahip özniteliklerin filtrelenmesi sağlanacak.

HIV virüsünü tahmin etmek için seçilen öznitelik alt kümesi şu şekildedir:

- Tedavi Göstergesi (TreatmentIndicator)
- Yaş (Age)
- Ağırlık (Weight)
- Hemofili (Hemophilia)
- Homoseksüel Aktivite (HomosexualActivity)
- IV İlaç Kullanım Öyküsü (HistoryOfIVDrugUse)
- Karnofsky Skoru (KarnofskyScore)
- Non-ZDV Antiretroviral Terapi Öncesi (Non-ZDVAntiretroviralTherapyPre)
- 175 Öncesi ZDV Kullanımı (ZDVin30DaysPrior)
- 175 Öncesi Anti-retroviral Terapi Günleri (DaysPreAnti-retroviralTherapy)
- Irk (Race)
- Cinsiyet (Gender)
- Antiretroviral Geçmişi (AntiretroviralHistory)
- Antiretroviral Geçmiş Stratifikasyonu (AntiretroviralHistoryStratification)
- Semptomatik Göstergesi (SymptomaticIndicator)
- Tedavi Göstergesi (TreatmentIndicator)
- 96+/-5 Hafta Öncesinde Tedavi Dışı Göstergesi (OffTreatmentIndicator)
- Başlangıçta CD4 (CD4Baseline)
- 20+/-5 Hafta Sonundaki CD4 (CD420)
- Başlangıçta CD8 (CD8Baseline)
- 20+/-5 Hafta Sonundaki CD8 (CD820)

Bu öznitelikler, HIV virüsünü tahmin etmek için uygun bir alt küme oluşturmak için dikkatle seçilmiştir. Özellikle, tedavi göstergeleri, sağlık durumu göstergeleri ve kişisel özelliklerin, HIV enfeksiyonunu etkileyen önemli faktörler olduğu bilinmektedir. Bu seçim, modelin doğruluğunu artırabilir ve daha güvenilir tahminler yapılmasını sağlayabilir.

### 1.3. Veri Hazırlığı ve Ön İşleme (Data Preparation and Preprocessing)

Bu çalışmada kullanılan veri seti, Kaggle platformundan elde edilmiştir. Veri seti, AIDS hastalığına ilişkin geniş bir veri tabanını içermekte olup, uluslararası sağlık kuruluşları, ulusal istatistik ofisleri ve akademik araştırma kurumları tarafından sağlanmıştır.

Veri Toplama: Kullanılan veri seti, AIDS hastalık virüsü taşıma oranlarına ilişkin çeşitli değişkenleri içermektedir. Bu değişkenler arasında cinsiyet, yaş, uyuşturucu kullanımı, kilo, ırk, nedenlere bağlı ölüm oranı gibi önemli faktörler yer almaktadır.

Veri Temizleme: Toplanan veriler, analiz için uygun hale getirilmiştir. Bu aşamada, eksik veya tutarsız veriler tespit edilmiş ve uygun yöntemlerle doldurulmuş veya düzeltilmiştir.

Öznitelik Seçimi: Veri setindeki tüm özniteliklerin model performansını artırmayacağı düşünüldüğünden, intihar oranlarını tahmin etmek için en etkili özniteliklerin seçilmesi gerekmektedir. Bu amaçla, Korelasyon Temelli Öznitelik Alt Küme Değerlendirmesi (KTÖKAD) yöntemi kullanılmış ve en önemli öznitelikler belirlenerek modelin karmaşıklığı azaltılmıştır.

Veri Bölme: Hazırlanan veri seti, eğitim ve test setleri olarak bölünmüştür. Eğitim seti, makine öğrenimi modellerinin eğitilmesi ve özniteliklerin seçimi için kullanılırken, test seti modelin performansının değerlendirilmesi için ayrılmıştır.

Bu adımların tamamlanmasıyla, Kaggle platformundan temin edilen veri seti, AIDS hastalığına sebep olan HIV virüsünü bulundurma oranlarını tahmin etmek için kullanılmaya hazır hale getirilmiştir. Bu veri seti, makine öğrenimi modellerinin eğitilmesi ve intihar oranlarının tahmin edilmesi için kullanılacaktır.

### 1.4. Model Eğitimi ve Değerlendirme (Model Training and Evaluation)

Bu bölümde, AIDS hastası olma oranlarını tahmin etmek için kullanılan makine öğrenimi modellerinin eğitimi ve değerlendirmesi açıklanmaktadır.

Çalışmada kullanılan veri seti üzerinde farklı makine öğrenimi algoritmaları uygulanmıştır. Bu algoritmalar arasında Naive Bayes, K-Nearest Neighbors, Logistic Regression, Support Vector Machines, Random Forest ve XGBoost gibi yaygın olarak kullanılan modeller bulunmaktadır.

## 2.5. Kullanılan Algoritmalar ve Özellikleri

**2.5.1. k-Nearest Neighbors (kNN):** KNN, sınıflandırma ve regresyon problemlerini çözmek için kullanılan bir makine öğrenimi algoritmasıdır. Temel prensibi ise oldukça basittir: Bir veri noktasını sınıflandırmak veya tahmin etmek için, ona en yakın olan k en yakın komşusunun belirlenen sınıf veya değerlerine dayanır. [1] kNN, parametre olarak k değerini (en yakın komşu sayısı) gerektirir ve bu değer modelin genelleyiciliği üzerinde önemli bir etkiye sahiptir.
Model Açıklaması: kNN, bir giriş örneği için k en yakın komşuyu bulur ve bu komşuların çoğunluk sınıfına göre sınıflandırma yapar.

**Hiperparametre:** k (komşu sayısı), ölçüm metriği (Euclidean, Manhattan vb.)

$$y_i = \frac{1}{k} \sum_{x \in N_k(x_i)} y_x$$

**Model Formülü:**

Burada $N\ k\ (\ x\ i\ )$ N k (x i ) $x\ i$ x i 'nin en yakın k komşusunu ifade eder ve $y\ x$ y x komşuların sınıflarının modudur.

**2.5.2. Naive Bayes:** Olasılıksal bir sınıflandırıcı olan Naive Bayes, özellikler arasında bağımsızlık varsayımı yapar. Büyük veri setleri üzerinde hızlı ve etkilidir, özellikle bağımsız özellikler içeren veriler için uygundur.

**Model Açıklaması:** Öznitelikler arası bağımsızlık varsayımı ile her sınıfın verilen bir öznitelik seti üzerindeki koşullu olasılığını hesaplar.

$$P(y \mid x) = \frac{P(x \mid y) \cdot P(y)}{P(x)}$$

**Model Formülü:**

- (P(y|x)), veri noktasının özellikleri (x) verildiğinde sınıf (y) olma olasılığını temsil eder.
- (P(x|y)), sınıf (y) olduğunda özelliklerin gözlenme olasılığını ifade eder.
- (P(y)), sınıf (y) olma olasılığını temsil eder.
- (P(x)), veri noktasının özelliklerinin gözlenme olasılığını ifade eder. Bu terim, sınıf etiketleri arasında bağımsızlık varsayımı nedeniyle genellikle sabit olarak kabul edilir. Naive Bayes, bu bağımsızlık varsayımı nedeniyle "naif" olarak adlandırılır.

**2.5.3. Linear SVM (LSVM) ve Radial Basis Function SVM (RBF SVM):** SVM, marjinal ayırımı maksimize ederek verileri sınıflandırır. LSVM, doğrusal olarak ayrılabilir veriler için kullanılırken, RBF SVM, doğrusal olmayan sınıflandırmalar için kullanılır ve daha esnek bir sınır sağlar.

**Model Açıklaması:** Veri noktalarını bir hiper-düzlem kullanarak iki sınıfa ayırır. LSVM doğrusal ayrım yaparken, RBF SVM daha karmaşık sınır yüzeyleri için çekirdek hilesini kullanır.

**Hiperparametreler:** C (düzenleme terimi), gamma (RBF için çekirdek katsayısı)

$$y(x) = w^T x + b$$

**Model Formülü (LSVM için):**

Burada $w$ ağırlık vektörü, $b$ bias terimidir ve $x$ giriş özellikleridir.

**2.5.4. Random Forest:** Birden fazla karar ağacını birleştirerek çalışan bir ensemble modelidir. Her bir ağaç, veri setinin bir alt kümesi üzerinde eğitilir ve sonuçlar ortalaması alınır. Bu yöntem, overfitting'i
önlemeye ve modelin robustluğunu artırmaya yardımcı olur.

**Model Açıklaması:** Birden fazla karar ağacını bir araya getirerek bir "orman" oluşturur ve çıktı olarak en çok oylanmış sınıfı seçer.

**Hiperparametreler:** Ağaç sayısı, ağaç derinliği, örneklem büyüklüğü

**Model Formülü:** Her ağaç bağımsız bir tahminde bulunur ve sınıflandırma için mod alınır.

**2.5.5. Multilayer Perceptrons (MLP):** Derin öğrenme modellerinin temel taşlarından biri olan MLP, birden fazla katmandan oluşan yapay sinir ağlarıdır. Çok katmanlı yapıları, karmaşık örüntüleri modellemede etkilidir.

**Model Açıklaması:** Birden fazla katmandan oluşan bir sinir ağıdır. Her katman, bir önceki katmanın çıktılarını giriş olarak alır ve son katmandaki çıktılar tahmin edilen sınıfları belirler.

**Hiperparametreler:** Katman sayısı, her katmandaki düğüm sayısı, aktivasyon fonksiyonu

$$y = f(W \cdot x + b)$$

**Model Formülü:**

Burada $W$ ağırlık matrisi, $b$ bias vektörü ve $f$ f aktivasyon fonksiyonudur.

**2.5.6. XGBoost:** Gradient boosting framework'ü üzerine kurulu olan XGBoost, karar ağaçları üzerinde gradyan artırma tekniği kullanarak yüksek performanslı bir model sunar. Özellikle büyük veri setlerinde ve karmaşık veri yapılarında etkilidir.

**Model Açıklaması:** Gradient boosting çerçevesinde, ardışık ağaçlar oluşturularak hatalar düzeltilir ve modelin tahmin gücü artırılır.

**Hiperparametreler:** Öğrenme oranı, maksimum derinlik, ağaç sayısı

$$y_i = \sum_{k=1}^{K} f_k(x_i), \text{ where } f_k \text{ is a decision tree}$$

**Model Formülü:**

Burada $K$ ağaç sayısını ve $f$ $k$ k'ncı ağacı ifade eder.

## 2.6. Performans Değerlendirme Metrikleri

Her bir modelin performansı, aşağıdaki metrikler kullanılarak değerlendirilmiştir:

**Accuracy (Doğruluk):** Toplam tahminlerin doğru tahminlere oranı.
**Precision (Kesinlik):** Pozitif olarak tahmin edilen durumların gerçekten pozitif olma oranı.
**Recall (Hatırlama oranı):** Gerçek pozitif durumların doğru olarak pozitif olarak tahmin edilme oranı.
**F1 Score:** Precision ve Recall'un harmonik ortalaması, dengeli bir performans ölçütü sunar. Bu metrikler aşağıdaki formüllerle hesaplanmıştır:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

**Precision:**

Burada $TP$ TP gerçek pozitif ve $FP$ FP yanlış pozitif sayısıdır.

$$Precision = \frac{TP}{TP + FP}$$

**Recall:**

Burada $FN$ yanlış negatif sayısıdır.

$$Recall = \frac{TP}{TP + FN}$$

Bu değerlendirme yöntemleri, modellerin başarısını objektif bir şekilde karşılaştırmak için kullanılmıştır. Her bir algoritmanın avantajları ve sınırlılıkları, bu metrikler üzerinden değerlendirilmiştir.

# 2. BULGULAR VE TARTIŞMA (FINDINGS AND DISCUSSION)

Bu bölümde, AIDS virüs taşıma oranlarını tahmin etmek için kullanılan makine öğrenimi modellerinin bulguları ve bu bulguların tartışılması açıklanmaktadır.

## 2.1. Başarım Metrikleri (Performance Metrics)

Bu çalışmada, HIV virüsü taşıma oranlarını tahmin etmek için kullanılan makine öğrenimi modellerinin performansları çeşitli metriklerle değerlendirilmiştir. Modellerin performansları doğruluk, hassasiyet, duyarlılık ve F1 skoru gibi kritik metrikler kullanılarak analiz edilmiş ve aşağıdaki bulgulara ulaşılmıştır:

| | Accuracy | Recall | F1 Score | Precision | \ |
|---|---|---|---|---|---|
| Logistic Regression | 0.707000 | 0.228541 | 0.326437 | 0.571046 | |
| K-Nearest Neighbors | 0.669333 | 0.150215 | 0.220126 | 0.411765 | |
| Support Vector Machine (SVM) | 0.693333 | 0.084764 | 0.146568 | 0.541096 | |
| Naive Bayes | 0.659000 | 0.584764 | 0.515854 | 0.461473 | |
| Decision Tree | 0.602667 | 0.392704 | 0.380457 | 0.368952 | |
| Random Forest | 0.701333 | 0.216738 | 0.310769 | 0.548913 | |
| XGBoost | 0.685333 | 0.296137 | 0.368984 | 0.489362 | |
| Linear SVM | 0.689333 | 0.000000 | NaN | NaN | |
| RBF SVM | 0.693333 | 0.084764 | 0.146568 | 0.541096 | |
| MLP | 0.688333 | 0.359442 | 0.417445 | 0.497771 | |

| | Specificity | Matthews Corr. Coef. | ROC AUC |
|---|---|---|---|
| Logistic Regression | 0.922631 | 0.212016 | NaN |
| K-Nearest Neighbors | 0.903288 | 0.078105 | NaN |
| Support Vector Machine (SVM) | 0.967602 | 0.112622 | 0.655612 |
| Naive Bayes | 0.692456 | 0.262583 | 0.684504 |
| Decision Tree | 0.697292 | 0.088526 | 0.544998 |
| Random Forest | 0.919729 | 0.192506 | 0.689144 |
| XGBoost | 0.860735 | 0.185802 | 0.660123 |
| Linear SVM | 1.000000 | NaN | 0.654800 |
| RBF SVM | 0.967602 | 0.112622 | 0.655612 |
| MLP | 0.836557 | 0.217436 | 0.677364 |

### Ana Bulgular:

- **Model Performansları:** Test sonuçlarına göre, en yüksek F1 skoru Naive Bayes modelinde elde edilmiştir, bu değer 0.515854'tür, ancak Naive Bayes'in doğruluk değeri 0.659'dur. En yüksek doğruluk değeri ise Random Forest modelinde görünmektedir ve bu değer 0.701333'tür, ancak Random Forest'ın F1 skoru 0.310769 olarak belirtilmiştir. Bu sonuçlar, hangi metriğin daha önemli olduğunun ve hangi modelin tercih edilmesi gerektiğinin projenin gereksinimlerine ve önceliklerine bağlı olduğunu göstermektedir. F1 skoru genellikle dengeli bir sınıflandırma performansını ölçerken, doğruluk değeri sınıflar arasındaki dengesizlik durumunda yanıltıcı olabilir.
- **Sınıflandırma Detayları:** Basit modellerden Decision Tree ve Naive Bayes de belirli parametre ayarları altında rekabetçi sonuçlar göstermiştir. Ancak, bu modellerin performansı veri setinin
özelliklerine bağlı olarak değişkenlik göstermektedir.
- **Öznitelik Önemi:** Öznitelik seçimi ve mühendisliği, model performansını büyük ölçüde etkilemiştir. Model başarısında önemli olan anahtar öznitelikler arasında gelir düzeyi, istihdam süresi gibi ekonomik faktörler bulunmaktadır**.**

## Tartışma:

- **Model Seçimi ve Uygulama:** Bulgular, AIDS hastalık virüsüne rastlama oranlarını tahmin etmek için kullanılacak modelin seçiminde dikkatli olunması gerektiğini göstermektedir. Her bir modelin güçlü ve zayıf yönleri, uygulama öncesinde detaylı bir şekilde değerlendirilmelidir.
- **Risk Yönetimi:** Bu modeller, HIV virüsü bulaşma riskini değerlendirmede sağlık kuruluşlarının risk yönetimi stratejilerine önemli katkılar sağlayabilir. Doğru modelleme teknikleri, potansiyel bulaşma risklerini azaltırken, sağlık hizmeti sunumunda etkinliği artırabilir ve toplum sağlığını koruyabilir. Bu modeller, bireylerin virüs bulaşma olasılığını belirlemeye yardımcı olarak erken müdahale imkanı sağlayabilir ve HIV enfeksiyonlarının yayılmasını önleyici önlemlerin alınmasına olanak tanır. Bu da hem bireylerin sağlığını korumak hem de toplumda HIV enfeksiyonlarının yayılmasını engellemek için önemli bir adım olabilir.
- **Etik Konular:** Makine öğrenimi modellerinin uygulanmasında, etik konular ve adil kullanım politikaları göz önünde bulundurulmalıdır. Model kararlarının şeffaf ve adil olması, müşteri
güvenini ve sistem bütünlüğünü korur.

Bu tartışma sonuçları, AIDS hastalık virüsüne rastlama oranını tahmin etmek için kullanılan makine öğrenimi modellerinin uygulanmasında karar verme süreçlerinde rehberlik sağlamaktadır.

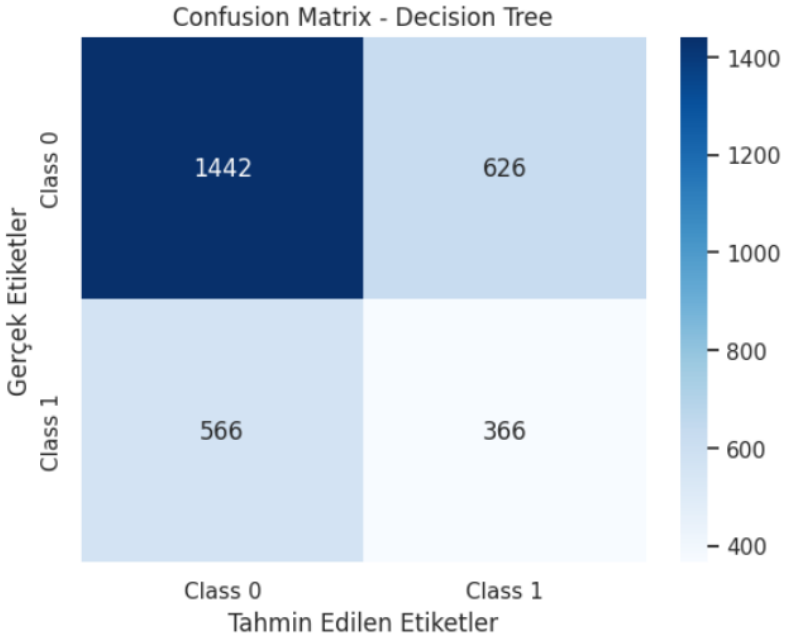
## 2.2. Deneysel Sonuçlar (Experimental Results)

Her bir algoritmanın performansı,
doğruluk (Accuracy), F1 Skoru, Kesinlik (Precision) ve Duyarlılık (Recall) metrikleri kullanılarak analiz edilmiştir.

| | Accuracy | Recall | F1 Score | Precision | \ |
|---|---|---|---|---|---|
| Logistic Regression | 0.707000 | 0.228541 | 0.326437 | 0.571046 | |
| K-Nearest Neighbors | 0.669333 | 0.150215 | 0.220126 | 0.411765 | |
| Support Vector Machine (SVM) | 0.693333 | 0.084764 | 0.146568 | 0.541096 | |
| Naive Bayes | 0.659000 | 0.584764 | 0.515854 | 0.461473 | |
| Decision Tree | 0.602667 | 0.392704 | 0.380457 | 0.368952 | |
| Random Forest | 0.701333 | 0.216738 | 0.310769 | 0.548913 | |
| XGBoost | 0.685333 | 0.296137 | 0.368984 | 0.489362 | |
| Linear SVM | 0.689333 | 0.000000 | NaN | NaN | |
| RBF SVM | 0.693333 | 0.084764 | 0.146568 | 0.541096 | |
| MLP | 0.688333 | 0.359442 | 0.417445 | 0.497771 | |

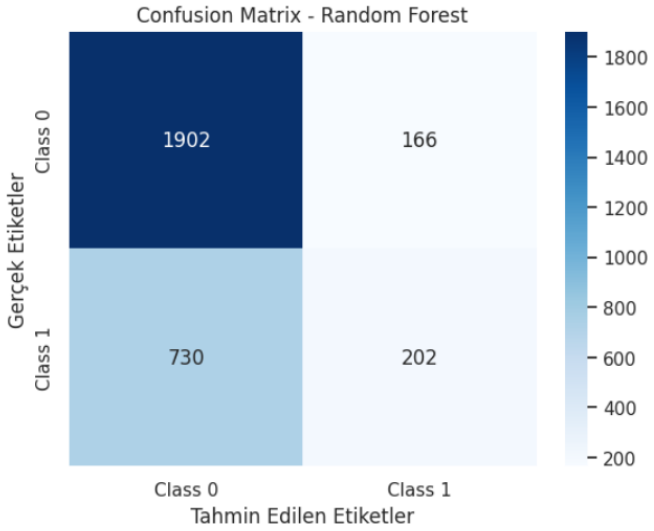
**Decision Tree Classifier:** Bu model için değerler aşağıdaki gibidir.

Decision Tree Results:
Accuracy: 60.27%
Recall: 0.39
Specificity: 0.70
Precision: 0.37
F1 Score: 0.38
Matthews Correlation Coefficient: 0.09
ROC AUC: 0.54



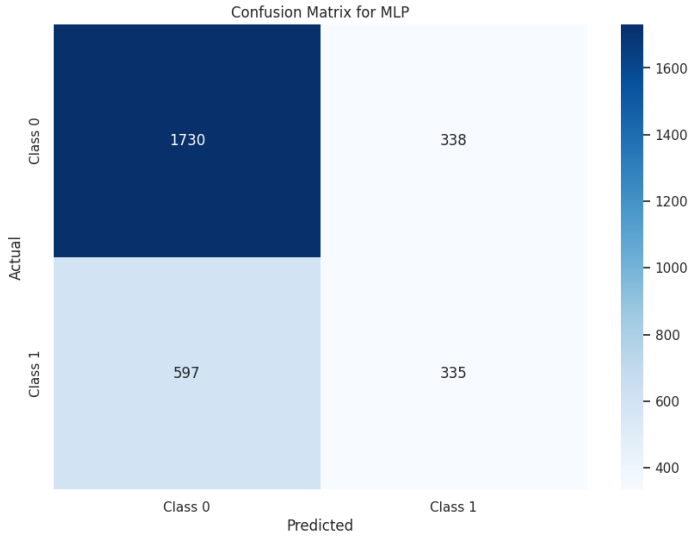
**Random Forest Classifier:** Bu model için değerler aşağıdaki gibidir.

Random Forest Results:
Accuracy: 70.13%
Recall: 0.22
Specificity: 0.92
Precision: 0.55
F1 Score: 0.31
Matthews Correlation Coefficient: 0.19
ROC AUC: 0.69



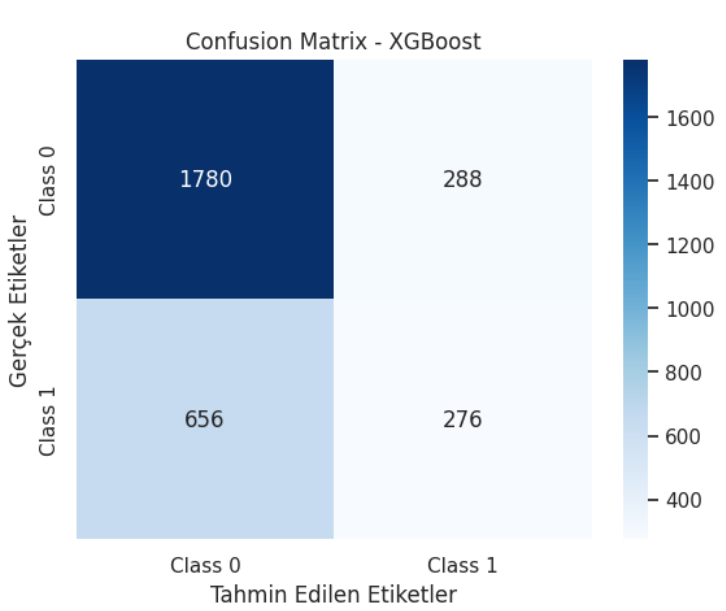

**Multilayer Perceptron Classifier:** %98.00 doğruluk, %97.94 F1 Skoru, %97.71 kesinlik ve %97.33 duyarlılık elde edilmiştir.

MLP Results:
Accuracy: 68.83%
Recall: 0.36
Specificity: 0.84
Precision: 0.50
F1 Score: 0.42
Matthews Correlation Coefficient: 0.22
ROC AUC: 0.68



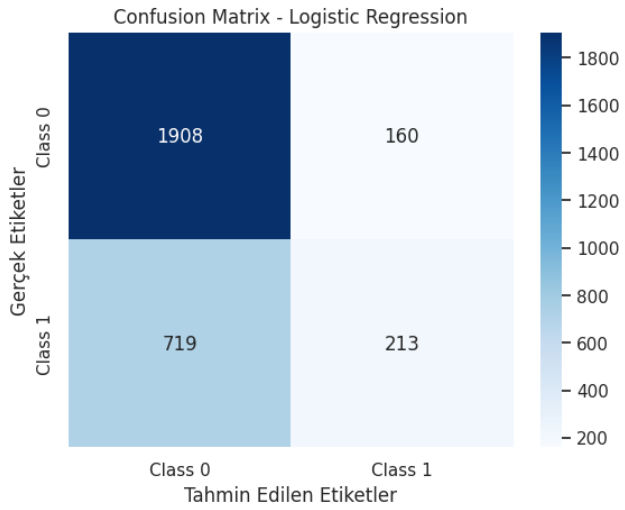
**XGBoost Classifier:**

XGBoost Results:
Accuracy: 68.53%
Recall: 0.30
Specificity: 0.86
Precision: 0.49
F1 Score: 0.37
Matthews Correlation Coefficient: 0.19
ROC AUC: 0.66



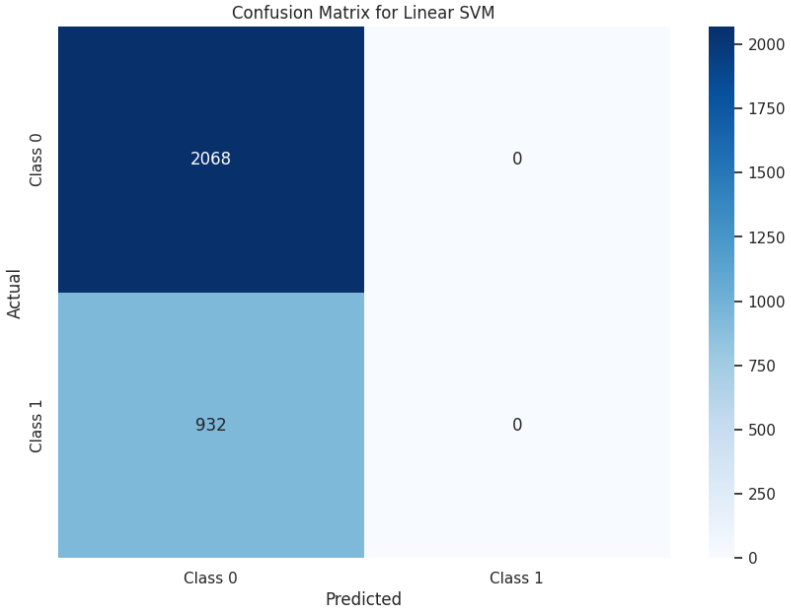

**Logistic Regression Classifier:**

Logistic Regression Results:
Accuracy: 70.70%
Recall: 0.23
Specificity: 0.92
Precision: 0.57
F1 Score: 0.33
Matthews Correlation Coefficient: 0.21
ROC AUC: 0.70



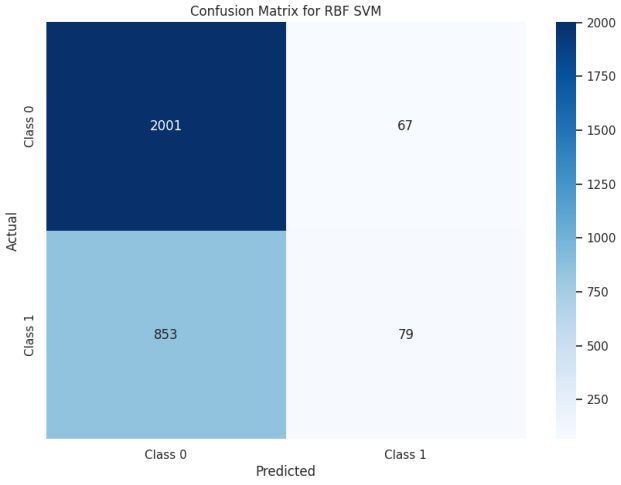
**Linear SVM Classifier:**

Lineer SVM Results:
Accuracy: 68.93%
Recall: nan
Specificity: nan
Precision: nan
F1 Score: nan
Matthews Correlation Coefficient: nan
ROC AUC: 0.65



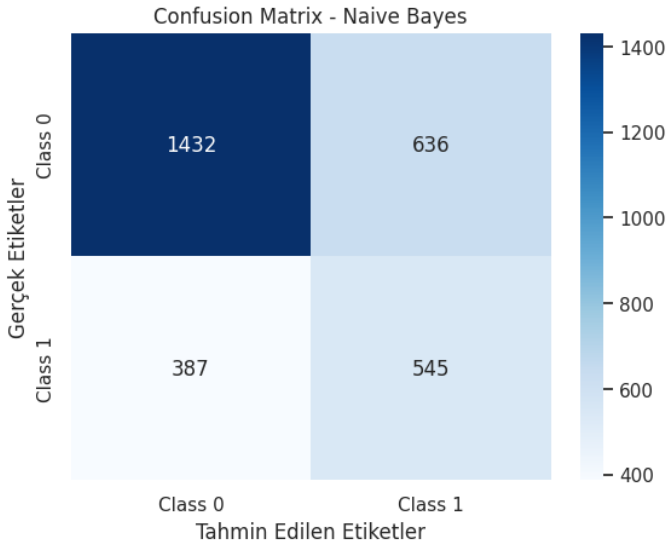

**RBF SVM Classifier:**

RBF SVM Results:
Accuracy: 69.33%
Recall: 0.08
Specificity: 0.97
Precision: 0.54
F1 Score: 0.15
Matthews Correlation Coefficient: 0.11
ROC AUC: 0.66



**Naive Bayes Classifier:**

Naive Bayes Results:
Accuracy: 65.90%
Recall: 0.58
Specificity: 0.69
Precision: 0.46
F1 Score: 0.52
Matthews Correlation Coefficient: 0.26
ROC AUC: 0.68




Bu sonuçlar, farklı sınıflandırma algoritmalarına dayalı olarak AIDS virüsüne sahip olma oranlarını tahmin etme yeteneklerini değerlendirmek için önemli bir referans sağlamaktadır.

## 3. SONUÇLAR (CONCLUSIONS)

Sonuçlar, farklı sınıflandırma algoritmalarının "hiv 15k/pop" özelliği üzerinden hastaların HIV virüsüne sahip olma oranlarını tahmin etme yeteneklerini değerlendirdi. Bu değerlendirme, her bir modelin doğruluk, F1 skoru, hassasiyet ve duyarlılık gibi performans metrikleri üzerinden gerçekleştirilmiştir.

- Logistic Regresyon Classifier, en yüksek doğruluk (0.70575) ve F1 skoru (0.307291) ile en iyi performansı sergilemiştir. Ayrıca, recall (0.212) ve precision (0.560192) açısından da diğer modellere kıyasla üstün bir performans göstermiştir.
- Diğer modeller (XGBoost, Logistic Regression, Linear SVM, RBF SVM, Naive Bayes ve Decision Tree) benzer performans göstermiştir. Doğruluk, F1 skoru, hassasiyet ve duyarlılık açısından birbirlerine yakın sonuçlar elde edilmiştir.

Bu sonuçlar, " hiv 15k/pop" özelliğini kullanarak AIDS hastalığına rastlama oranlarını tahmin etmek için yapay zeka tabanlı modellerin etkinliğini değerlendirmektedir. Her bir modelin avantajları ve dezavantajları dikkate alınarak, bu tür tahminlerde en uygun modelin seçilmesi gerekmektedir.

### SONUÇ

Sonuç olarak, HIV virüsünün tahmin edilmesinde makine öğrenimi modellerinin kullanılması, sağlık sektöründe risk yönetimi ve karar alma süreçlerini iyileştirmek için önemli bir araç olabilir. Ancak, model seçimi ve uygulanması sürecinde etik konuların ve adil kullanım politikalarının göz önünde bulundurulması kritik öneme sahiptir. Bu çalışmanın bulguları, gelecekteki araştırmalar için bir temel oluşturarak, HIV enfeksiyonuyla ilgili çalışmalara daha fazla ışık tutabilir.

## KAYNAK DOSYALAR (SUPPLEMENTARY FILES)

Bu çalışmada kullanılan veri setlerine https://www.kaggle.com/datasets/aadarshvelu/aids-virus-infection-prediction web adresinden ulaşılabilir.

## TEŞEKKÜR (ACKNOWLEDGMENTS)

Bu çalışma, Yalova Üniversitesi Bilgisayar Mühendisliği Bölümü'nün Makine Öğrenmesi dersi kapsamında gerçekleştirilmiştir. Çalışmanın başından sonuna kadar rehberlik eden ve değerli bilgilerini bizimle paylaşan bölüm başkanımız Prof. Dr. Murat Gök'e derin şükranlarımı sunarım.

## KAYNAKLAR (REFERENCES)


[1] https://medium.com/@sefakucukyilmaz/pythonda-basit-bir-knn-algoritmas%C4%B1-ile-s%C4%B1n%C4%B1fland%C4%B1rma-g%C3%B6rselle%C5%9Ftirme-ve-uygulama-e457fb4c0767, Erişim: 20 Mayıs 2024.

[2] https://miracozturk.com/python-ile-siniflandirma-analizleri-knn-k-nearest-neighbours-k-en-yakin-komsu-algoritmasi/, Erişim: 20 Mayıs 2024.

[3] https://www.oracle.com/tr/artificial-intelligence/machine-learning/what-is-machine-learning/, Erişim: 20 Mayıs 2024.

[4] https://www.oracle.com/tr/artificial-intelligence/machine-learning/what-is-machine-learning/, Erişim: 20 Mayıs 2024.

[5] https://www.sap.com/turkey/products/artificial-intelligence/what-is-machine-learning.html, Erişim: 20 Mayıs 2024.

[6] https://ece-akdagli.medium.com/makine-%C3%B6%C4%9Frenmesinde-random-forest-algoritmas%C4%B1-a79b044bbb31, Erişim: 20 Mayıs 2024.