**T.C.**

**SAKARYA ÜNİVERSİTESİ**

**Bilgisayar ve Bilişim Bilimleri Fakültesi**

**Bilgisayar Mühendisliği Bölümü**

**AĞ GÜVENLİĞİ**

# Network Security ile İlgili Bilinen Veri Setleri Üzerinde Araştırma ve Makine Öğrenme Algoritmalarının Performans Analizi

**Hazırlayan**
**İkranur Ayça Gecü**
**B201210094**
**Rabia Nur Çağlı**
**B201210350**
**1A**

# GİRİŞ

Ağ güvenliği, günümüzde internetin yaygın kullanımıyla birlikte giderek artan bir öneme sahip olmuştur. İnternet üzerindeki veri trafiğinin hızla artması ve çeşitlenmesi, aynı zamanda çeşitli siber tehditlerin ortaya çıkmasına neden olmuştur. Bu tehditler arasında veri sızıntıları, bilgi hırsızlıkları ve bilgisayar ağlarına yönelik saldırılar gibi unsurlar, ağ güvenliği uzmanlarını daha sofistike ve etkili güvenlik önlemleri geliştirmeye yönlendirmiştir.

Bu bağlamda, network security (ağ güvenliği) alanındaki çalışmalar, bu alanla ilgili veri setleri ile oluşturulan modeller ile saldırı tespit sistemi geliştirilmesi üzerine odaklanmaktadır. Veri setleri, çeşitli ağ trafiği senaryolarını simüle ederek, gerçek dünya koşullarında ağ güvenliği önlemlerinin etkinliğini değerlendirmek adına değerli bir kaynak teşkil etmektedir.

Bu çalışmanın temel amacı, ağ güvenliği ile ilgili veri setleri üzerinde farklı modellerin oluşturulması ve bu modellerin performans analizini gerçekleştirmektir. Bu analiz, farklı makine öğrenmesi tekniklerinin kullanıldığı durumları içermektedir. Elde edilen sonuçlar, ağ güvenliği uzmanlarının, mevcut tehditlere karşı daha etkili koruma stratejileri geliştirmelerine katkı sağlamayı amaçlamaktadır.

## NETWORK SECURITY İLE İLGİLİ VERİ SETLERİ

**1)** KDD Cup 1999 Veri Seti

KDD Cup'99 veri seti Saldırı tespit sistemlerinin testi için geliştirilmiş ve 5. Uluslararası Knowledge Discovery and Data Mining konferansı çerçevesinde düzenlenen bir yarışmada kullanıma sunulan veri setidir. Bu veri setindeki saldırılar 4 ana kategoride sınıflandırılabilir.

- Denial of Service (Hizmet Engelleme): Bu saldırılar genel olarak TCP/IP protokol yapısındaki açıklardan faydalanılarak bir sunucuya birden çok bağlantı isteği göndererek yasal kullanıcıların hizmet almasını engellemeye yöneliktir.
- Bilgi Tarama (Probing): Bu tür saldırılar bir sunucunun yada herhangi bir makinanın geçerli IP adreslerini, aktif portlarını veya işletim sistemini öğrenmek için geliştirilmiştir.
- Yönetici Hesabı ile Yerel Oturum Açma (Remote to LocalR2L): Kullanıcı haklarına sahip olunmadığı durumda misafir yada başka bir kullanıcı olarak izinsiz erişim yapılmasıdır.
- Kullanıcı Hesabının Yönetici Hesabına Yükseltilmesi (User to RootU2R): Bu tip saldırılarda sisteme girme izni olan fakat yönetici olmayan bir kullanıcının yönetici izni gerektirecek işler yapmaya kalkımasıdır. KDD Cup'99 veri seti hem eğitim hem de test verisini içermektedir. Eğitim verisinde toplam 494020 örnek mevcuttur. Eğitim ve Test setindeki her bir örnek toplam 41 özellikten oluşmaktadır. Bu özellikler ise Basit, İçerik, Zaman Tabanlı Trafik ve Host Tabanlı Trafik adı altında 4 farklı grupta toplanmıştır.

**2)** UNSW-NB15 Veri Seti

UNSW-NB15 veri seti IXIA PerfectStorm aracı kullanılarak Avustralya siber güvenlik merkezi laboratuvarlarında hem gerçek modern normal aktivite hem de yapay ,günümüz şartlarına uygun ağ trafiği saldırı hareketlerini içeren hibrit bir model oluşturulmuştur. Tcpdump aracı ile 100 GB işlenmemiş ağ trafiğini yakalanmış ve ARgus ve Bro-IDS vb. araçlar 12 model veri setindeki özellikleri çıkarmak için geliştirilmiştir (Moustafa and Slay 2016). Veri setinin geliştiricileri ayrıca veri setini eğitim veri seti ve test veri seti olarak iki farklı gruba da ayırmıştır. Bu veri seti Avrupa Bilim ve Teknoloji Dergisi e ISSN: 2148-2683 110 daha sonra birçok araştırmacı tarafından da kullanılmıştır (Moustafa and Slay 2016). Eğitim veri seti 175,341 kayıttan, test veri seti 82,332 kayıttan oluşmaktadır. Orijinal veri seti ise 2,540,044 kayıttan oluşmaktadır. (Sonule et al. 2020). Eğitim ve test veri setinin saldırı sınıflarına göre dağılımları:

| Sınıf | Eğitim Seti Seti | Test |
|---|---|---|
| Normal | 56,000 | 37,000 |
| Analysis | 2,000 | 677 |
| Backdoor | 1,746 | 583 |
| DoS | 12,264 | 4,089 |
| Exploits | 33,393 | 11,132 |
| Fuzzers | 18,184 | 6,062 |
| Generic | 40,000 | 18,871 |
| Reconnaissance | 10,491 | 3,496 |
| Shellcode | 1,133 | 378 |
| Worms | 130 | 44 |
| Toplam Kayıt Sayısı | 175,341 | 82,332 |

UNSW-NB15 veri seti toplam 49 özellik ve 1 hedef değere sahiptir.

**3)** CICIDS 2017 Dataset

CICIDS 2017 (Canadian Institute for Cybersecurity Intrusion Detection Systems 2017), siber güvenlik alanında kullanılmak üzere özel olarak tasarlanmış bir veri setidir. Bu veri seti, ağ güvenliği ve saldırı tespiti sistemlerinin geliştirilmesi amacıyla çeşitli ağ tabanlı saldırı senaryolarını simüle etmek üzere oluşturulmuştur.

Geniş bir özellik yelpazesi sunan kapsamlı bir veri setidir. Bu özellikler arasında ağ trafiği, paket başlıkları, protokol istatistikleri, kullanıcı davranışları ve benzeri ağ aktivitelerine dair detaylı bilgiler bulunmaktadır. Veri seti, normal ağ trafiği yanında çeşitli siber saldırıları da içermektedir, bunlar arasında DDoS saldırıları, kötü amaçlı yazılımlar, botnet aktiviteleri gibi farklı senaryolar yer almaktadır.

CICIDS 2017, ağ güvenliği uzmanlarına, araştırmacılara ve öğrencilere gerçekçi bir ortamda çalışma imkanı sunmayı hedeflemektedir. Bu veri seti üzerinde yapılan çalışmalar, siber tehditlere karşı daha etkili ve güvenilir saldırı tespiti sistemlerinin geliştirilmesine katkıda bulunmayı amaçlamaktadır.

Bu veri setinin sütun bilgisi aşağıda verilmiştir.

| No. | Feature Name | No. | Feature Name | No. | Feature Name |
|---|---|---|---|---|---|
| 1 | Flow ID | 29 | Fwd IAT Std | 57 | ECE Flag Count |
| 2 | Source IP | 30 | Fwd IAT Max | 58 | Down/Up Ratio |
| 3 | Source Port | 31 | Fwd IAT Min | 59 | Average Packet Size |
| 4 | Destination IP | 32 | Bwd IAT Total | 60 | AvgFwd Segment Size |
| 5 | Destination Port | 33 | Bwd IAT Mean | 61 | AvgBwd Segment Size |
| 6 | Protocol | 34 | Bwd IAT Std | 62 | FwdAvg Bytes/Bulk |
| 7 | Time stamp | 35 | Bwd IAT Max | 63 | FwdAvg Packets/Bulk |
| 8 | Flow Duration | 36 | Bwd IAT Min | 64 | FwdAvg Bulk Rate |
| 9 | Total Fwd Packets | 37 | Fwd PSH Flags | 65 | BwdAvg Bytes/Bulk |
| 10 | Total Backward Packets | 38 | Bwd PSH Flags | 66 | BwdAvg Packets/Bulk |
| 11 | Total Length of FwdPck | 39 | Fwd URG Flags | 67 | BwdAvg Bulk Rate |

**MAKİNE ÖĞRENMESİ ALGORİTMALARI**

Ağ güvenliği alanında kullanılan temel makine öğrenimi algoritmaları şunlardır:

1) Karar Ağaçları (Decision Trees)
    - Ağaç yapısında karar verme kurallarını temsil eden bir modeldir.
    - Ağ güvenliği için sızma tespiti, zararlı yazılım algılama ve trafiği sınıflandırma gibi alanlarda kullanılabilir.
    - Basit ve açıklayıcıdır, ancak tek bir ağaç genellikle aşırı uyuma eğilimlidir.

2) Destek Vektör Makineleri (Support Vector Machines - SVM)
    - İki sınıf arasında karar sınırlarını belirleyen ve maksimum marjini (mesafeyi) en üst düzeye çıkarmaya çalışan bir algoritmadır.
    - Ağ güvenliğinde sınıflandırma, sızma tespiti ve saldırı algılama için kullanılabilir.
    - Karmaşık ilişkileri yakalayabilir ve aşırı uyuma karşı dirençlidir.

3) Rastgele Ormanlar (Random Forests)
    - Birden fazla karar ağacını birleştirerek daha güçlü ve genelleştirilebilir bir model oluşturan bir ensemble (topluluk) algoritmasıdır.
    - Sızma tespiti, saldırı sınıflandırması ve trafiği analizi gibi alanlarda etkili olabilir.

4) Doğrusal ve Lojistik Regresyon:
    - Doğrusal regresyon, bir bağımlı değişkenin bir veya daha fazla bağımsız değişkenle ilişkisini modellemek için kullanılır.
    - Lojistik regresyon, sınıflandırma problemleri için kullanılır. Özellikle binary (iki sınıflı) sınıflandırmalarda etkilidir.

5) K-En Yakın Komşu (KNN)
    - Bir veri noktasını sınıflandırmak veya tahmin yapmak için komşularını kullanır.
    - Ağ trafiği sınıflandırma, saldırı tespiti ve normal trafiği ayırt etme gibi alanlarda kullanılabilir.

6) Naive Bayes
    - Bayes teoremine dayanan olasılık temelli bir sınıflandırma algoritmasıdır.
    - E-posta spam tespiti, zararlı yazılım algılama ve sızma tespiti gibi alanlarda kullanılabilir.

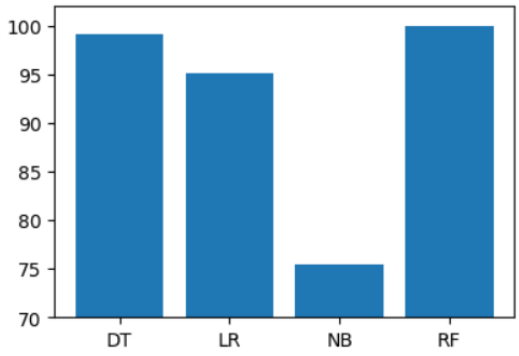
7) Yapay Sinir Ağları (Artificial Neural Networks - ANN)
    - Biyolojik sinir sistemlerinden ilham alarak oluşturulan ve çok katmanlı yapıları olan karmaşık modellerdir.
    - Ağ güvenliğinde genellikle derin öğrenme modelleri olarak kullanılır; sızma tespiti, zararlı yazılım algılama ve trafiği analizi gibi alanlarda etkilidir.

## PERFORMANS ANALİZLERİ

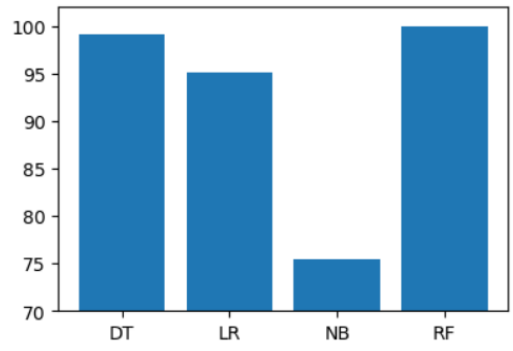
### CICIDS 2017 Dataseti kullanılarak makine öğrenme yöntemlerinin performans analizi

Bu çalışmada kullanılan sınıflandırma algoritmaları Random Forest , Decision Tree, Naive Bayes ve Lineer Regresyon 'dur. Bu modeller eğitildikten sonra alınan performans çıktıları şunlardır:

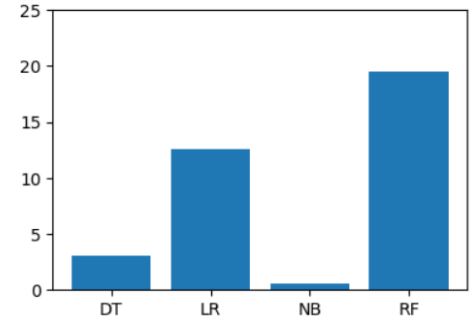
## TRAINING ACCURACY



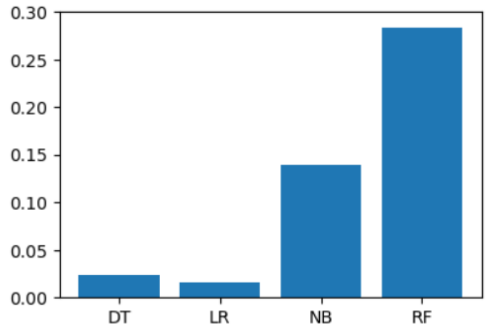

## TEST ACCURACY



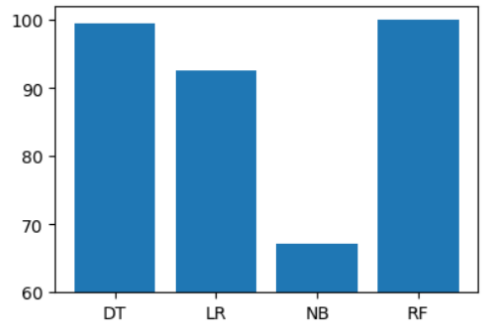

## TRAINING TIME



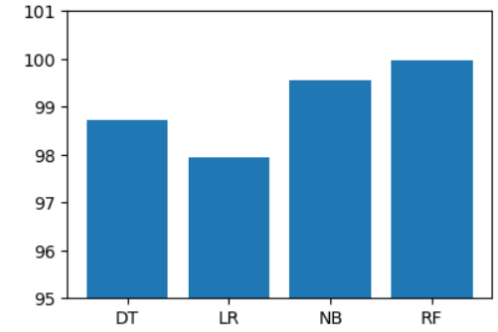

## TESTING TIME



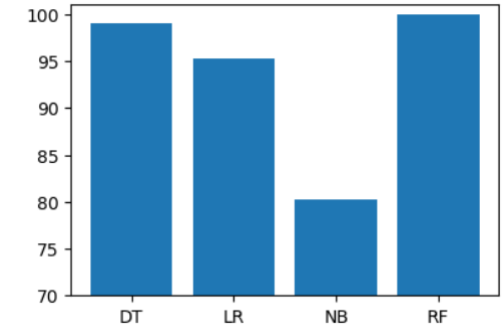

## PRECISION ON TRAINING SET




## RECALL ON TRAINING SET




## F1 SCORE ON TRAINING SET



## 1. Model Performansı

### 1.1 Training ve Test Accuracy

**Decision Tree (DT):**

- Training Accuracy: 99.0838%
- Test Accuracy: 99.0689%

**Linear Regression (LR):**

- Training Accuracy: 95.0827%
- Test Accuracy: 95.1116%

**Naive Bayes (NB):**

- Training Accuracy: 75.4274%
- Test Accuracy: 75.4274%

**Random Forest (RF):**

- Training Accuracy: 99.9865%
- Test Accuracy: 99.9745%

### 1.2 Precision, Recall ve F1 Score

**Decision Tree (DT):**

- Training Precision: 99.4304%
- Training Recall: 98.7327%
- Training F1 Score: 99.0804%

**Linear Regression (LR):**

- Training Precision: 92.6498%
- Training Recall: 97.9322%
- Training F1 Score: 95.2178%

**Naive Bayes (NB):**

- Training Precision: 67.1471%
- Training Recall: 99.54903%
- Training F1 Score: 80.1989%

**Random Forest (RF):**

- Training Precision: 99.9980%
- Training Recall: 99.97502%
- Training F1 Score: 99.9865%

## 2. Eğitim ve Test Süreleri

**Decision Tree (DT):**

- Eğitim Süresi: 3.016832 saniye
- Test Süresi: 0.024205 saniye

**Linear Regression (LR):**

- Eğitim Süresi: 12.535989 saniye
- Test Süresi: 0.016349 saniye

**Naive Bayes (NB):**

- Eğitim Süresi: 0.599255 saniye
- Test Süresi: 0.139145 saniye

**Random Forest (RF):**

- Eğitim Süresi: 19.520218 saniye
- Test Süresi: 0.284000 saniye

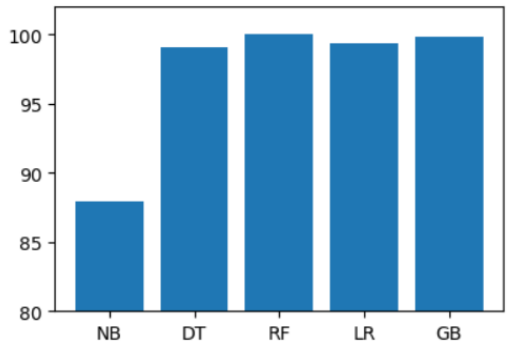
## 3. Sonuçlar ve Değerlendirme

- Model Performansı: Random Forest (RF) hem eğitim hem de test setlerinde en yüksek doğruluk oranına sahip modeldir (%99.98).
- Eğitim Süreleri: Naive Bayes (NB ) en hızlı eğitilen modeldir (0.5992 saniye) ancak test süresi açısından en hızlı model Linear Regression (LR) modelidir (0.016 saniye).
- Precision ve Recall: Naive Bayes (NB), diğer modellere kıyasla yüksek recall değerine sahiptir ancak düşük precision ile dengelenmemiş bir modeldir.
- F1 Score: Random Forest (RF) en yüksek F1 score değerine sahip olan modeldir, yani hem precision hem de recall açısından dengeli bir performans sergilemektedir.

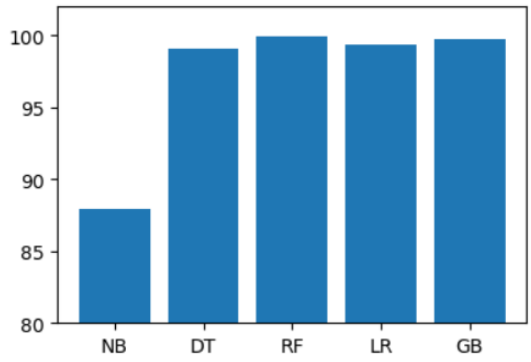
**NSL-KDD Dataseti kullanılarak makine öğrenme yöntemlerinin performans analizi**

Bu çalışmada kullanılan sınıflandırma algoritmaları Random Forest , Decision Tree, Naive Bayes ve Lineer Regresyon ve Gradient Booting Classifier'dır. Bu modeller eğitildikten sonra alınan performans çıktıları şunlardır:

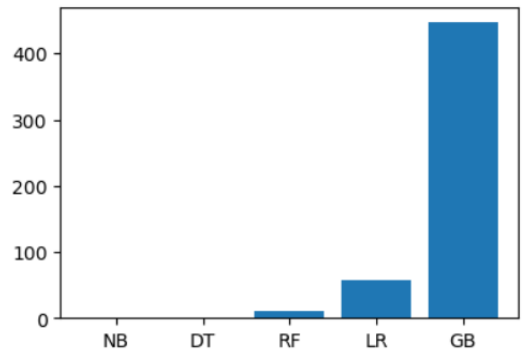
**TRAINING ACCURACY**



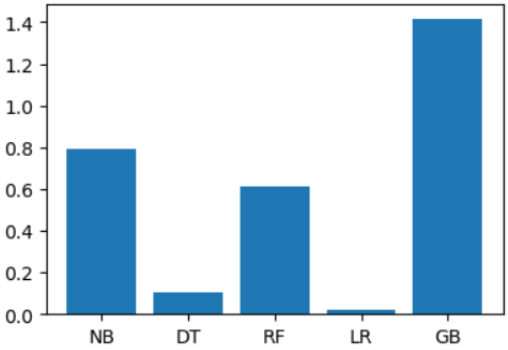

**TEST ACCURACY**



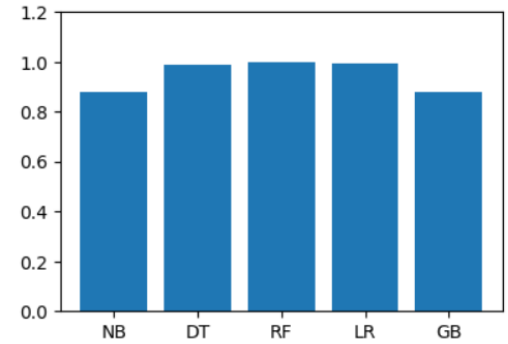

**TRAINING TIME**



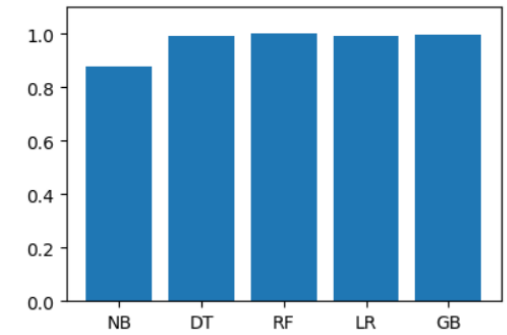

**TESTING TIME**



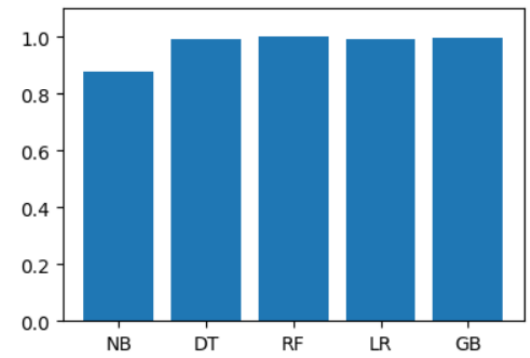

**PRECISION ON TRAINING SET**




**RECALL ON TRAINING SET**




**F1 SCORE ON TRAINING SET**



**1. Model Performansı**

**1.1 Training ve Test Accuracy**

**Naive Bayes (NB):**

- Training Accuracy: 87.951%
- Test Accuracy: 87.903%

**Decision Tree (DT):**

- Training Accuracy: 99.05%
- Test Accuracy: 99.052%

**Random Forest (RF):**

- Training Accuracy: 99.997%
- Test Accuracy: 99.969%

**Linear Regression (LR):**

- Training Accuracy: 99.352%
- Test Accuracy: 99.352%

**Gradient Booting Classifier**

- Training Accuracy: 99.793%
- Test Accuracy: 99.771%

**1.2 Precision, Recall ve F1 Score**

**Naive Bayes (NB):**

- Training Precision: 0.879038
- Training Recall: 0.879038441
- Training F1 Score: 0.879038441

**Decision Tree (DT):**

- Training Precision: 0.990523
- Training Recall: 0.99052304
- Training F1 Score: 0.990523042

**Random Forest (RF):**

- Training Precision: 0.99968103
- Training Recall: 0.9996810344
- Training F1 Score: 0.999681034

**Linear Regression (LR):**

- Training Precision: 0.99352867
- Training Recall: 0.9935286792
- Training F1 Score: 0.99352867929

**Gradient Booting Classifier**

- Training Precision: 0.87903844
- Training Recall: 0.99771816938
- Training F1 Score: 0.9977181693

**2. Eğitim ve Test Süreleri**

**Naive Bayes (NB):**

- Eğitim Süresi: 1.04721 saniye
- Test Süresi: 0.79089 saniye

**Decision Tree (DT):**

- Eğitim Süresi: 1.50483 saniye
- Test Süresi: 0.10471 saniye

**Random Forest (RF):**

- Eğitim Süresi: 11.45332 saniye
- Test Süresi 0.60961 saniye

**Linear Regression (LR):**

- Eğitim Süresi: 56.67286 saniye
- Test Süresi: 0.02198 saniye

**Gradient Booting Classifier**

- Eğitim Süresi: 446.69099 saniye
- Test Süresi: 1.41416 saniye

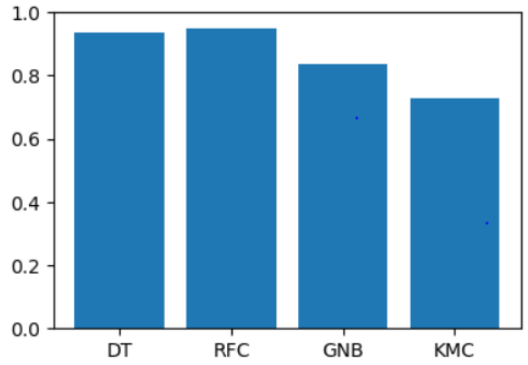
**3. Sonuçlar ve Değerlendirme**

- Model Performansı: Random Forest (RF) hem eğitim hem de test setlerinde en yüksek doğruluk oranına sahip modeldir (%99.9..).
- Eğitim Süreleri: Naive Bayes (NB ) en hızlı eğitilen modeldir ancak test süresi açısından en hızlı model Linear Regression (LR) modelidir
- Precision ve Recall: Random Forest, diğer modellere kıyasla yüksek precision ve recall değerine sahiptir.
- F1 Score: Random Forest (RF) en yüksek F1 score değerine sahip olan modeldir, yani hem precision hem de recall açısından dengeli bir performans sergilemektedir.

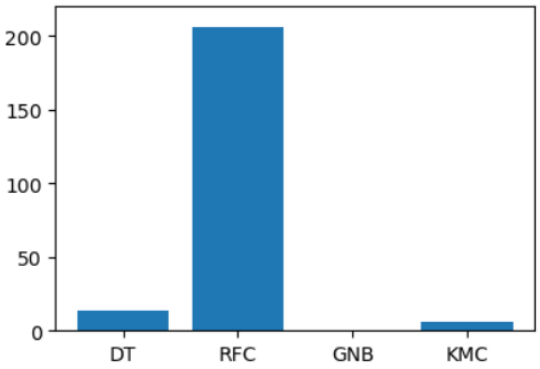
## UNSW_NB15 Dataseti kullanılarak makine öğrenme yöntemlerinin performans analizi

Bu çalışmada kullanılan sınıflandırma Decision Tree Classifier, Random Forest Classifier, Gaussian Naive Bayes ve K-Means Clusteringdir. Bu modeller eğitildikten sonra alınan performans çıktıları şunlardır:

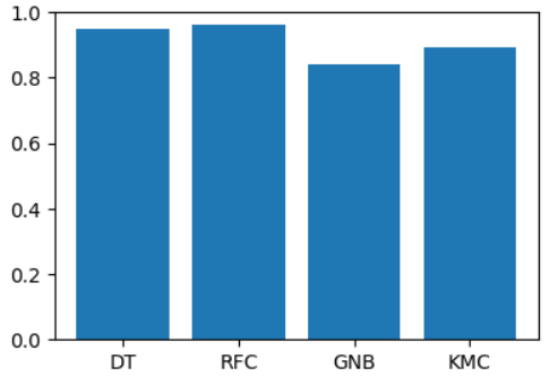
### TEST ACCURACY



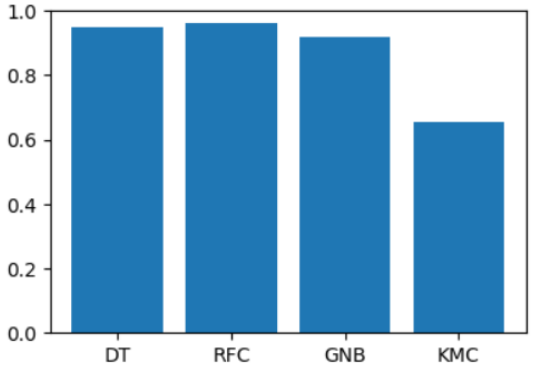

### TRAINING TIME



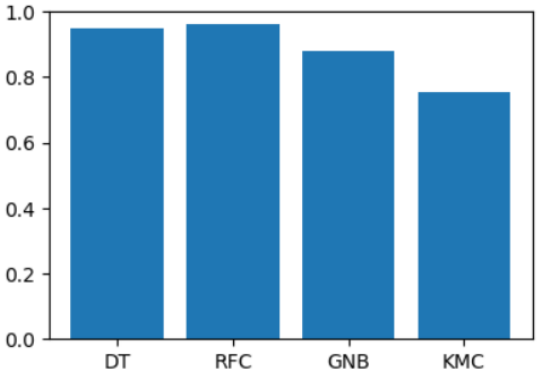

### PRECISION ON TRAINING SET




### RECALL ON TRAINING SET




### F1 SCORE ON TRAINING SET




| | Accuracy | Precision | Recall | Training time | f1 score |
|---|---|---|---|---|---|
| Decision Tree Classifier | 0.936561 | 0.950842 | 0.949958 | 13.359766 | 0.950400 |
| Random Forest Classifier | 0.951295 | 0.963634 | 0.960108 | 206.077450 | 0.961868 |
| Gaussian Naive Bayes | 0.837404 | 0.841875 | 0.918355 | 0.379936 | 0.878454 |
| K-Means Clustering | 0.728429 | 0.894023 | 0.652938 | 5.679137 | 0.754694 |

**1. Model Performansı**

**1.1 Test Accuracy**

**Decision Tree (DT):**

- Test Accuracy: 0.936561

**Random Forest Classifier (RFC):**

- Test Accuracy: 0.951295

**Gaussian Naive Bayes (GNB):**

- Test Accuracy: 0.837404

**K-Means Clustering (KMC):**

- Test Accuracy: 0.728429

**1.2 Precision, Recall ve F1 Score**

**Decision Tree (DT):**

- Training Precision: 0.950842
- Training Recall: 0.949958
- Training F1 Score: 0.950400

**Random Forest Classifier (RFC):**

- Training Precision: 0.963634
- Training Recall: 0.960108
- Training F1 Score: 0.961868

**Gaussian Naive Bayes (GNB):**

- Training Precision: 0.841875
- Training Recall: 0.918355
- Training F1 Score: 0.878454

**K-Means Clustering (KMC):**

- Training Precision: 0.894023
- Training Recall: 0.652938
- Training F1 Score: 0.754694

**2. Eğitim Süreleri**

**Decision Tree (DT):**

- Eğitim Süresi: 13.359766

**Random Forest Classifier (RFC):**

- Eğitim Süresi: 206.077450

**Gaussian Naive Bayes (GNB):**

- Eğitim Süresi: 0.379936

**K-Means Clustering (KMC):**

- Eğitim Süresi: 5.679137

**3. Sonuçlar ve Değerlendirme**

- Model Performansı: Random Forest (RF) en yüksek doğruluk oranına sahip modeldir (0.951295).
- Eğitim Süreleri: Gaussian Naive Bayes (NB ) en hızlı eğitilen modeldir (0.379936 saniye
- Precision ve Recall: Gaussian Naive Bayes (NB), diğer modellere kıyasla en yüksek precision değerine de recall değerine de sahip olan modeldir..
- F1 Score: Random Forest (RF) en yüksek F1 score değerine sahip olan modeldir..