**Zararlı Yazılım Tespiti Raporu**

**Angelo Oliveira, 7 Kasım 2019, "Malware Analysis Datasets: Top-1000 PE Imports", IEEE Dataport, doi:** <https://dx.doi.org/10.21227/004e-v304> // https://www.kaggle.com/datasets/ang3loliveira/malware-analysis-datasets-top1000-pe-imports

**1. Giriş**

Bu projenin amacı, zararlı ve zararsız yazılım örneklerini makine öğrenme tekniklerini kullanarak ayırt edebilen bir model geliştirmektir. Veri seti, çeşitli yazılım örneklerinden çıkarılan en iyi 1000 Portable Executable (PE) import özelliklerini içerir ve bu örnekler zararlı (1) veya zararsız (0) olarak etiketlenmiştir.

**2. Veri Seti Genel Bakış**

Veri seti, her bir yazılım örneği için benzersiz bir hash kimliği ve zararlı olup olmadığını belirten ikili bir etiketin yanı sıra 1000 PE import özelliğini içermektedir.

**3. Özellik Seçimi**

Model performansını artırmak ve hesaplama maliyetini azaltmak için ExtraTreesClassifier kullanılarak özellik seçimi yapılmıştır. SelectFromModel yöntemi ile en önemli özellikler seçilmiştir.

**4. Veri Bölme**

Veri seti, %70 eğitim ve %30 test oranında bölünmüştür.

**5. Model Eğitimi ve Değerlendirilmesi**

Seçilen özellikler kullanılarak çeşitli makine öğrenme modelleri eğitilmiş ve değerlendirilmiştir. Bu modeller arasında K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machine (SVM), Decision Tree, Random Forest ve Logistic Regression bulunmaktadır. Derin öğrenme Algoritmalarıda ise SNN. CNN, RNN, GNA ve GRU bulunmaktadır. Her modelin sınıflandırma performansı precision, recall ve F1-score ile değerlendirilmiştir.

**5.1.1 K-Nearest Neighbors (KNN)**

**Performans:**

* **Precision**: 0.97
* **Recall**: 0.97
* **F1-Score**: 0.97

**5.1.2 Naive Bayes**

**Performans:**

* **Precision**: 0.85
* **Recall**: 0.85
* **F1-Score**: 0.85

**5.1.3 Support Vector Machine (SVM)**

**Performans:**

* **Precision**: 0.99
* **Recall**: 0.99
* **F1-Score**: 0.99

**5.1.4 Decision Tree**

**Performans:**

* **Precision**: 0.98
* **Recall**: 0.98
* **F1-Score**: 0.98

**5.1.5 Random Forest**

**Performans:**

* **Precision**: 0.99
* **Recall**: 0.99
* **F1-Score**: 0.99

**5.1.6 Logistic Regression**

* **Precision**: 0.95
* **Recall**: 0.95
* **F1-Score**: 0.95

**5.2.1 SNN**

**Performans:**

* **Precision**: 0.98
* **Recall**: 1.00
* **F1-Score**: 0.99

**5.2.2 CNN**

**Performans:**

* **Precision**: 0.98
* **Recall**: 1.00
* **F1-Score**: 0.99

**5.2.3 RNN**

**Performans:**

* **Precision**: 0.96
* **Recall**: 0.99
* **F1-Score**: 0.98

**5.2.4 GNA**

**Performans:**

* **Precision**: 0.97
* **Recall**: 0.98
* **F1-Score**: 0.97

**5.2.5 GRU**

**Performans:**

* **Precision**: 0.96
* **Recall**: 0.99
* **F1-Score**: 0.98

**5.3 Performans Karşılaştırması ve Değerlendirme**

Bu çalışmada kullanılan modellerin performans metrikleri karşılaştırıldığında, özellikle Basit Yapay Sinir Ağı (SNN) ve Convolutional Neural Network (CNN) modellerinin en yüksek performansı gösterdiği görülmektedir. Her iki model de precision, recall ve f1-score metriklerinde 0.98 ve üzeri değerler alarak oldukça yüksek bir doğruluk sağlamaktadır.

**5.3.1 SNN ve CNN Karşılaştırması**

SNN ve CNN modelleri, her ikisi de 0.98 precision, 1.00 recall ve 0.99 f1-score değerlerine ulaşmıştır. Bu sonuçlar, her iki modelin de zararlı yazılım tespitinde çok başarılı olduğunu göstermektedir. SNN'nin daha basit yapısı, hızlı eğitim ve tahmin süreci sağlarken, CNN'nin daha karmaşık yapısı, derin öğrenme katmanları sayesinde daha ayrıntılı öğrenme kapasitesi sunar.

**5.3.2 RNN ve GRU Karşılaştırması**

RNN ve GRU modelleri, precision ve recall metriklerinde oldukça benzer sonuçlar göstermiştir. Her iki model de 0.96 precision ve 0.99 recall değerlerine ulaşmıştır. Ancak, GRU'nun f1-score'u 0.98 ile biraz daha yüksekken, RNN'nin f1-score'u da aynıdır. Bu sonuçlar, her iki modelin de zaman serisi verilerini işleme kabiliyetlerinin güçlü olduğunu göstermektedir.

**5.3.3 GNA Performansı**

GNA modeli, precision, recall ve f1-score metriklerinde diğer modellere göre biraz daha düşük performans göstermiştir. Precision ve recall değerleri sırasıyla 0.97 ve 0.98 olup, f1-score'u 0.97'dir. Bu durum, GNA modelinin diğer modellere göre biraz daha az başarılı olduğunu göstermektedir.

**6. Sonuç**

Bu çalışmada değerlendirilen modeller arasında, SNN ve CNN en yüksek performansı gösterirken, RNN ve GRU modelleri de oldukça başarılı sonuçlar elde etmiştir. GNA modeli ise diğer modellere göre biraz daha düşük performans göstermiştir.

Sonuç olarak, zararlı yazılım tespiti için hem basit hem de karmaşık yapay sinir ağı modelleri etkili bir şekilde kullanılabilir. Ancak, model seçimi yapılırken eğitim ve tahmin süresi gibi faktörler de göz önünde bulundurulmalıdır. SNN ve CNN modelleri, yüksek doğruluk oranları ile zararlı yazılım tespiti için en uygun modeller olarak öne çıkmaktadır.