# COM3551: ARTIFICIAL INTELLIGENCE

Malware Detection using Machine Learning

Bedirhan Alp Arslan 21290662
13.08.23

**Introduction:**

This project's aim is to use machine learning to analyze PE imports, which offer insights about the behavior of binary executables.

In recent years the number of cyber attacks have increased exponentially costing people and organizations a substantial amount of resources.

The traditional solutions can’t keep up with the amount of new threats every day. Predicting if an executable is malicious or not would help reduce the costs of cyber attacks and increase peoples digital safety.

**Methodology:**

1. **Dataset and Preprocessing:** The dataset comprises the 1000 PE imports and is loaded using Pandas
2. **Feature Selection:** Extra Trees Classifier is utilized to identify influential features within the dataset. The most important attributes are selected for subsequent analysis.
3. **Model Training:** The dataset is divided into training and testing sets. KNN, NB, and SVM are chosen as classification algorithms for their simplicity and effectiveness in classification tasks.
4. **Model Evaluation:** Classification reports are generated for each algorithm, providing precision, recall, and F1-score metrics. These metrics assess the models' ability to accurately classify malware and benign samples.

**Results:**

1. **Feature Selection:** Extra Trees Classifier identifies significant features from PE imports, enabling a reduction in feature dimensionality. The model identified 222 features among the initial 1000 features.
2. **Model Performance:**
    - **KNN:** Achieves satisfactory precision and F1-score, but lower recall, indicating a balance between accuracy and comprehensive detection. Accuracy is %98.
    - **NB:** Demonstrates competitive performance across all metrics, highlighting its robustness for malware classification. Accuracy is %92.
    - **SVM:** Displays similar results to KNN, with a trade-off between precision and recall. Accuracy is %98.

**Conclusion:**

This study underscores the potential of machine learning for malware analysis using PE imports. By employing KNN, NB, and SVM, the project showcases trade-offs between precision and recall in classifying malware samples. The results emphasize the viability of these algorithms for automated malware detection. Further exploration could involve using more advanced models for enhanced accuracy and generalization, implementing cross validation and using a more balanced dataset.

**Code & Dataset**

https://www.kaggle.com/balpars/malware-detection

Angelo Oliveira, November 7, 2019, "Malware Analysis Datasets: Top-1000 PE Imports", IEEE Dataport, doi: https://dx.doi.org/10.21227/004e-v304.