**METHODOLOGY PROPOSED PE FILE MALWARE CLASSIFICATION SYSTEM**

**Dataset:-**

[**https://www.kaggle.com/datasets/amauricio/pe-files-malwares**](https://www.kaggle.com/datasets/amauricio/pe-files-malwares)

**About Dataset**

**Context**

This dataset is a result of a my research about Machine Learning & Malware Detection. It was built using a Python Library and contains benign and malicious data from PE Files. Below are the feature extracted from exe file using pefile library.

['e\_magic', 'e\_cblp', 'e\_cp', 'e\_crlc', 'e\_cparhdr', 'e\_minalloc', 'e\_maxalloc', 'e\_ss', 'e\_sp', 'e\_csum', 'e\_ip', 'e\_cs', 'e\_lfarlc', 'e\_ovno', 'e\_oemid', 'e\_oeminfo', 'e\_lfanew', 'Machine', 'NumberOfSections', 'TimeDateStamp', 'PointerToSymbolTable', 'NumberOfSymbols', 'SizeOfOptionalHeader', 'Characteristics', 'Magic', 'MajorLinkerVersion', 'MinorLinkerVersion', 'SizeOfCode', 'SizeOfInitializedData', 'SizeOfUninitializedData', 'AddressOfEntryPoint', 'BaseOfCode', 'ImageBase', 'SectionAlignment',

'FileAlignment', 'MajorOperatingSystemVersion', 'MinorOperatingSystemVersion', 'MajorImageVersion', 'MinorImageVersion', 'MajorSubsystemVersion', 'MinorSubsystemVersion', 'SizeOfHeaders', 'CheckSum', 'SizeOfImage', 'Subsystem', 'DllCharacteristics', 'SizeOfStackReserve', 'SizeOfStackCommit', 'SizeOfHeapReserve', 'SizeOfHeapCommit', 'LoaderFlags', 'NumberOfRvaAndSizes']

**Inspiration**

Machine Learning Antimalware

**Tools & Libraries used:-**

Python , colab ,VS Code, pandas, numpy, scikit-learn, matplotlib, seaborn, plotly,flask etc.

**Project Flow :-**

**1 Importing all libraries :-** import all required libraries which is used for implementation.

**2 Data Loading :-** loading csv file into pandas data frame.

**3 Data Cleaning:-** cleaning data includes null values handling, outlier removal, removing unnecessary columns etc.

**4 Data Visualization :-** creating plot and graph for better understanding of dataset and finding insight from it.

**5 Data Preprocessing :-** Label encoding (converting categorical to numbers, smote over sampling for balancing data etc.

**6 Feature importance :-** finding top impact full feature on target column using model feature selection technique .

**7 Splitting data :-** into train and test .

**8 Model Building** (model mentioned below).

**9 Model Evaluation :-** Confusion matrix, Classification Report .

**10 Web Application (Flask):-** input exe file , extracting feature from exe file using pefile library, and model prediction .

**MACHINE LEARNING MODELS**

1. ADA BOOST CLASSIFIER
2. GRADIENT BOOSTING CLASSIFIER
3. RANDOM FOREST CLASSIFIER
4. STACKING CLASSIFIER [ classifiers = RandomForestClassifier(), GradientBoostingClassifier()],

meta\_classifier = AdaBoostClassifier() ]