**Machine Learning Planning**

1. **Dataset:**

Existing Ones:

* [API Call based Malware Dataset](https://www.kaggle.com/datasets/focatak/malapi2019)
* [Windows\_Malware\_Detection\_Dataset](https://www.kaggle.com/datasets/radwahashiesh/windows-malware-detection-dataset?rvi=1)
* [VHS-22](https://www.kaggle.com/datasets/f3a18ca1e8f028aee166b12128e9532dc96d1005f3eb33e18db92c145939bc55)
* [SDN Dataset for DDoS Flooding Attack Detection](https://www.kaggle.com/datasets/mobayodeakinsolu/sdn-dataset-for-ddos-flooding-attack-scenarios)
* [UCI](https://archive.ics.uci.edu/datasets?skip=0&take=10&sort=desc&orderBy=NumHits&search=malware)
* <https://www.archive.ics.uci.edu/dataset/855/tuandromd+(tezpur+university+android+malware+dataset)>
* [Malware Executable Detection Dataset](https://www.kaggle.com/datasets/piyushrumao/malware-executable-detection/data)

More:

<https://www.unb.ca/cic/datasets/index.html>

<https://registry.opendata.aws/sorel-20m/>

Create Alone or to make new ones when the ML is Working : <https://cuckoo.sh/blog/>

1. **Model:**

Random Forests:

Type: Ensemble Learning

Description: Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. It is robust, handles high-dimensional data well, and is less prone to overfitting.

1. **General Description Of What The Program Will Do**

File will run in a sandbox and will write is behavior then use the data set of good ware and malware to decide if the file may have malicious intentions or not.

Or  
It will to determent what files may acting maliciously and what know and with a small data about is behavior will decide.