Introduction: The Microsoft Malware Prediction System is a trailblazing solution in a digital environment with risks. This method, which uses machine learning to forecast and prevent malware assaults, promises to raise cybersecurity standards for individuals and enterprises. Malware is a pervasive threat that targets computer systems worldwide in various forms, including viruses, ransomware, and spyware. Microsoft's solution to this problem combines cutting-edge algorithms and enormous databases to identify developing threats, adjust, and safeguard people in real time. Its key strength is the Microsoft Malware Prediction System's capacity to spot trends and anomalies and anticipate new threats. It provides a level of protection that is unmatched and equips users to protect their digital assets, ensuring a secure computer environment.

Motivation: With Microsoft Windows OS dominating 87% of the desktop OS market, the urgency to safeguard against malware attacks is paramount. To address this challenge, Microsoft initiated a call to action for the data science community. Our mission is to develop machine learning systems capable of predicting and preventing malware infections on Windows systems.

The motivation behind this endeavor is straightforward to protect the data and security of millions of users. Malware attacks are a growing concern, and this project is a proactive step towards fortifying defenses. By creating a system that predicts the likelihood of a machine being infected, my aim to empower users with enhanced security and preserve their digital integrity. This initiative reflects Microsoft's unwavering commitment to providing robust, innovative solutions for a safer digital world.

Dataset Description: For this problem, I collected data i.e train.csv and test.csv directly from kaggle where this competition was hosted.

Train.csv:- This file contains many entries where each entry corresponds to a machine which is uniquely identified by a MachineIdentifier and HasDetections is the ground truth and indicates that Malware was detected on the machine. And contains 83 columns including MachineIdentifier and HasDetections using which model is to be trained. Here features, columns and variables are analogous with each other and I will be using this terms interchangeably throughout this blog.

Test.csv:- This file contains 82 columns including MachineIdentifier except HasDetections.