Taxonomie des attaques sur les méthodes d'apprentissage automatique

Tom Djaaleb*, Mohamed-Lamine Messai*

*Univ Lyon, Univ Lyon 2, ERIC
{tom.djaaleb, mohamed-lamine.messai}@univ-lyon2.fr

Résumé. L'apprentissage automatique gagne de plus en plus de terrains d'applications. Différentes méthodes existent et qui permettent la construction de modèles à des fins d'aide à la décision. Néanmoins, les modèles d'apprentissage automatique sont vulnérables et exposés à différents types d'attaques de sécurité durant le processus d'apprentissage des modèles et après leur déploiement. Par conséquent, ces menaces doivent être, dans un premier temps, identifiées, définies et classées afin que, dans un deuxième temps, proposer des mesures de défense pour faire face à ces menaces. Dans cet article, Nous nous sommes intéressés à l'étude des menaces pouvant toucher un processus d'apprentissage automatique. Nous présentons une classification des menaces autour de l'objectif, la connaissance et la capacité de l'attaquant. Ensuite, nous montrons quelques exemples d'attaques sur des applications utilisant l'apprentissage automatique.

1 Introduction

L'apprentissage automatique ou machine (*Machine Learning (ML)*) est un sous-domaine de l'intelligence artificielle. Le ML inclut un ensemble d'algorithmes permettant de créer automatiquement des modèles à partir de données exemples. Ces données, utilisées dans la phase d'entraînement sont appelées les jeux de données d'apprentissage. Traditionnellement, un programme informatique effectue une tâche en suivant des instructions précises, et donc systématiquement de la même façon. Par contre, un système d'apprentissage automatique ne suit pas d'instruction, mais apprend à partir de l'expérience. Par conséquent, ses performances s'améliorent au fil de son entraînement sur le jeu de données. La finalité du modèle de ML est d'imiter un comportement semblable au cerveau humain en termes de prise de décision afin d'automatiser des tâches. Ces dernières années, grâce à des capacités de calcul beaucoup plus puissantes et des ensembles de données beaucoup plus importants pour l'entraînement des modèles, les technologies d'apprentissage automatique, en particulier les réseaux neuronaux artificiels et les architectures d'apprentissage profond, ont fait des progrès considérables. L'apprentissage automatique a de nombreux domaines d'application comme la détection des spams et des logiciels malveillants en cybersécurité, la classification d'images et la reconnaissance d'objets pour le contrôle des véhicules autonomes, le diagnostic médical, la reconnaissance vocale, ... etc.

Avec le succès impressionnant de l'application du ML dans des domaines de plus en plus nombreux, des faiblesses de sécurité sont présentes dans les étapes des algorithmes d'appren-