# Medical Text Classification Using Supervised Machine Learning

Ce projet s'intéresse à l'utilisation d'algorithmes d'apprentissage supervisé pour la classification automatique de textes médicaux, dans le but de faciliter l'analyse et l'interprétation de données médicales souvent complexes. Pour y parvenir, plusieurs méthodes de représentation du texte (TF-IDF, Word2Vec, GloVe, SBERT et BioBERT) ont été intégrées à des algorithmes de classification courants, parmi lesquels Naive Bayes, les arbres de décision, la régression logistique et les machines à vecteurs de support (SVM). Les performances de ces différentes approches ont été évaluées à l'aide des indicateurs de précision, rappel, F1-score et support.

Les expérimentations réalisées mettent en évidence l'intérêt particulier des modèles SBERT et BioBERT pour la compréhension du vocabulaire médical spécialisé. En effet, ces modèles bénéficient d'un apprentissage approfondi sur des corpus biomédicaux, offrant ainsi de meilleurs résultats que les approches plus traditionnelles comme TF-IDF ou GloVe, surtout en contexte de petits ensembles de données. Parmi les algorithmes de classification testés, les SVM se révèlent particulièrement performants grâce à leur capacité à gérer des représentations de haute dimension et à s'adapter aux subtilités du langage médical.

Au-delà de la recherche de performances optimales, ce travail souligne également des défis importants, notamment la nécessité de disposer de données variées et suffisamment nombreuses, ou encore l'importance de mécanismes d'adaptation pour traiter des termes rares ou des abréviations spécifiques au domaine médical. Les résultats obtenus contribuent à l'avancée de l'informatique médicale en fournissant des orientations concrètes pour les travaux futurs et des pistes d'intégration pratique de systèmes d'analyse textuelle automatisée en milieu clinique.