Evento	Salão UFRGS 2022: SIC - XXXIV SALÃO DE INICIAÇÃO CIENTÍFICA DA UFRGS
Ano	2022
Local	Campus Centro - UFRGS
Título	Identificação de biomarcadores em dados genômicos com aprendizado de máquina
Autor	LUCAS LIMA DE MELO
Orientador	MARIANA RECAMONDE MENDOZA GUERREIRO

Uma das principais dificuldades no estudo de dados genômicos é a capacidade de encontrar genes capazes de ajudar a compreender e tratar doenças complexas como câncer. Modelos de aprendizado de máquina (AM) são comumente utilizados para tentar resolver este problema devido a sua capacidade de lidar com o grande volume e complexidade destes dados. Este estudo tem como objetivo identificar potenciais biomarcadores estáveis em problemas de diagnóstico ou prognóstico de câncer a partir de dados genômicos utilizando modelos de AM ensemble, ou seja, a partir da combinação de múltiplos algoritmos. Foram coletados dados de metilação de DNA de câncer de tireoide no GEO, os quais foram submetidos a pré-processamento para normalização dos dados. Uma metodologia para seleção de atributos baseada em ensemble heterogêneo foi implementada utilizando quatro métodos de seleção de atributos (T-Test,F-Test,Mutual Information e ReliefF), os quais foram agregados utilizando o algoritmo borda count para gerar o resultado do ensemble. Três algoritmos de classificação (SVM,Naive Bayes e LightGBM) foram usados para analisar o poder preditivo dos genes selecionados pelos métodos de seleção de atributos através de uma validação cruzada. Este processo de validação foi selecionado para diminuir o viés de avaliação dos modelos. Em um total de 20 iterações, utilizando a seleção de atributos ensemble heterogênea, o modelo obteve média de 93% na métrica ROC AUC score usando apenas cinco sondas como biomarcadores para predição, em um conjunto de dados que contém mais de vinte mil atributos originalmente.Nossos resultados indicam bom poder de generalização e boa capacidade de diagnóstico utilizando a metodologia aplicada,devido a alta sensibilidade do modelo, uma das métricas essenciais e mais desejadas no campo da medicina de precisão.