Metodologia do Experimento

Após efetuar levantamento dos principais métodos de seleção de características e tipos de dados utilizados no prognóstico do câncer de mama. Serão utilizadas bases de dados públicas que contenham os tipos de dados clínicos do paciente e dados moleculares da neoplasia analisada, para obter qual o desempenho dos modelos de aprendizagem de máquina aplicados neste contexto.

Os participantes do objeto de estudo serão os disponíveis na base de dados pública utilizada no experimento. Para obter os *datasets* necessários para o experimento, serão utilizadas plataformas que dispõem de dados que podem ser obtidos de forma gratuitas, por exemplo, a plataforma *Kagle*. Esta pode ser definida como a maior plataforma de hospedagem de projetos e competições de *Data Science*, ela disponibiliza kernels, datasets, além de dispor de um fórum de perguntas.

Após concluir a etapa de aquisição dos dados a serem utilizados no experimento, estes devem passar por um processo de tratamento, para que possam ser utilizados nas técnicas de seleção de características, e que não comprometam o desempenho dos modelos de aprendizagem de máquina. Nesse processo de tratamento deve-se, por exemplo, eliminar instâncias com dados faltantes, bem como remover instâncias duplicadas.
Além disso, deve-se levar em conta as considerações éticas sobre os dados utilizados, onde estes não devem conter nenhuma informação pessoal dos pacientes cujos dados estão inseridos nos datasets analisados.

Por fim, serão utilizadas algumas métricas como *acurácia*, *sensibilidade*, *precisão*, *revocação* e *curva RoC*, para avaliar o

desempenho dos modelos de aprendizagem de máquina utilizados sobre as características oriundas das técnicas de seleção de características usadas no experimento.