Neste trabalho, um algoritmo de *Machine Learning*, o *Random Forest*, será utilizado para gerar duas classificações da cidade do Rio de Janeiro. Serão classificações binárias, que terão como unidade os setores censitários do IBGE (via censo de 2022). Cada setor, portanto, será classificado como favela ou não-favela. O desempenho de cada uma das classificações será comparado. A referência serão os polígonos de favela do IBGE utilizados no censo de 2022.

A cidade do Rio de Janeiro foi escolhida para análise uma vez que possui grandes complexos de favelas, como o da Rocinha e o da Maré, nos quais atuam diversas redes de pesquisa, fornecendo uma gama de informações adicionais. O termo “favela”, cuja origem remonta à Guerra de Canudos, na Bahia, está historicamente vinculado também à formação urbana carioca, uma vez que seus ex-combatentes, desabonados pelo Estado, instauraram-se e fundaram o Morro da Providência, no Rio de Janeiro. Ainda, a capital do estado possui em números absolutos a segunda maior população residente em favelas do Brasil, perdendo apenas para a cidade de São Paulo, maior metrópole do país (IBGE, 2022).

O trabalho inicia-se pela análise exploratória de dados espaciais, com o objetivo de testar a hipótese nula de aleatoriedade espacial na distribuição de indicadores socioeconômicos (como renda, esgoto, lixo, arborização) em relação a áreas de favela. Para isso, será calculado o Índice Global de Moran, verificando se há autocorrelação espacial significativa que justifique a rejeição da aleatoriedade. Confirmada a existência de padrões espaciais (clusters), a pesquisa avançará para a construção de um modelo, utilizando a Matriz de Proximidade para gerar variáveis de contexto espacial.

Na etapa de modelagem, serão treinados **dois** classificadores *Random Forest*. Um deles será alimentado pelos indicadores de renda, esgoto, lixo, arborização e afins do IBGE, e outro será alimentado por esses e por variáveis derivadas da estrutura espacial, como a Média Espacial Móvel e Indicadores Locais de Associação Espacial (LISA), especificamente o Índice Local de Moran, como *features* categóricas, permitindo ao algoritmo distinguir diferentes regimes espaciais, como agrupamentos de valores similares ou *outliers* locais.

Os modelos serão comparados para ver qual atingiu o melhor desempenho e a sua interpretação será realizada via SHAP (*Shapley Additive Explanations*), buscando decifrar a “black box” dos algoritmos de *Machine Learning*.