Four datasets were tested in this project. The first dataset includes all of the features provided. The second set contains only microbiome-related features. The third set considers only patients with M. avium or M. abcesssus infections. Finally, the fourth set includes all features but contains only 2 samples per patient. Prior to modelling, missing values were imputed and set to the average value for that feature based on available data.

I used the LIBSVM (library for support vector machines) software package to calculate F-scores and to create and test predictive models. I used a slightly modified version of the script ‘fselect.py’, which is available from the authors of LIBSVM, to calculate F-scores for potential features and then, based on those results, to test many combinations of those features as potential models. The potential models are tested by a five-fold cross validation process. The complete F-score and modelling results are reported in Feature\_ranking\_results\_v5.xlsx.