# Report

Objectives

The main objective is to build a model that can correctly predict in-campus locations using Wi-Fi signal strength from various routers around the buildings. This indoors positioning proves to be tricky as the multiple devices interfere with each other and prevents geo-location using the usual GPS signals.

Data Cleaning and optimization

Initially I tried to run the raw data through the entire modelling algorithm of Random Forest, with the computing power I have, it was able to run over a period of 6hrs. However, this had to be improved for the other algorithms, therefore, I took advantage of R parallelization capabilities and used the *doParallel* library to utilize 7 cores out of the 8 available on my machine. Furthermore, I reduced the number of instances from 19,937 on each of the 529 features to only 5249 instances over the 529 features. This was achieved by only looking at one building (building 0). This combination allowed my computer to runt the random forest modeling algorithm in 5 minutes. With these new settings in place I ran 3 other algorithms: K-Nearest Neighbors (KNN), SVM Radial, and C50.

Results

Random Forest – All data results

Text

Description automatically generated

Random Forest - Building 0

Text

Description automatically generated

C50 - Building 0

Graphical user interface, text

Description automatically generated

SVM Radial Fit - Building 0

Text

Description automatically generated

K-Nearest Neighbors (KNN) – Building 0

Text

Description automatically generated

Summary of Results

A black screen with white text

Description automatically generated with low confidence

Based on these results, the best algorithm to use is the C50 algorithm, however if run time is not an issue, the Random Forest prediction using the full range of data is the best choice with the highest Accuracy and Kappa scores.

Recommendations

Attempting other algorithms with using cloud computing services from amazon (aws) is an option that is worth mentioning. For this project I attempted to use these but proved to be more problematic and complicated than just cleaning and refining the data.

Taking principal components analysis (PCA) and applying it to this large amount of data can be a useful tool to further increase the time output for each model.