Technical Report

Deep Analytics and Visualization

Evaluate Techniques for Wi-Fi Locationing

Greg Hepguler

BUSINESS OBJECTIVES

- Develop a system that will help people navigate large, complex, unfamiliar interior spaces, such as buildings in large industrial campuses, business complex, commercial property or shopping malls without getting lost

- Investigate feasibility of using Wi-Fi fingerprinting to benchmark activity at certain locations and compare that activity to other locations to assess location-based marketing

- Select a relevant dataset to perform data analytics, representative of required location-based solutions for malls, convention centers or large business complex

- Recommend the best model that produces best metrics for dependable results

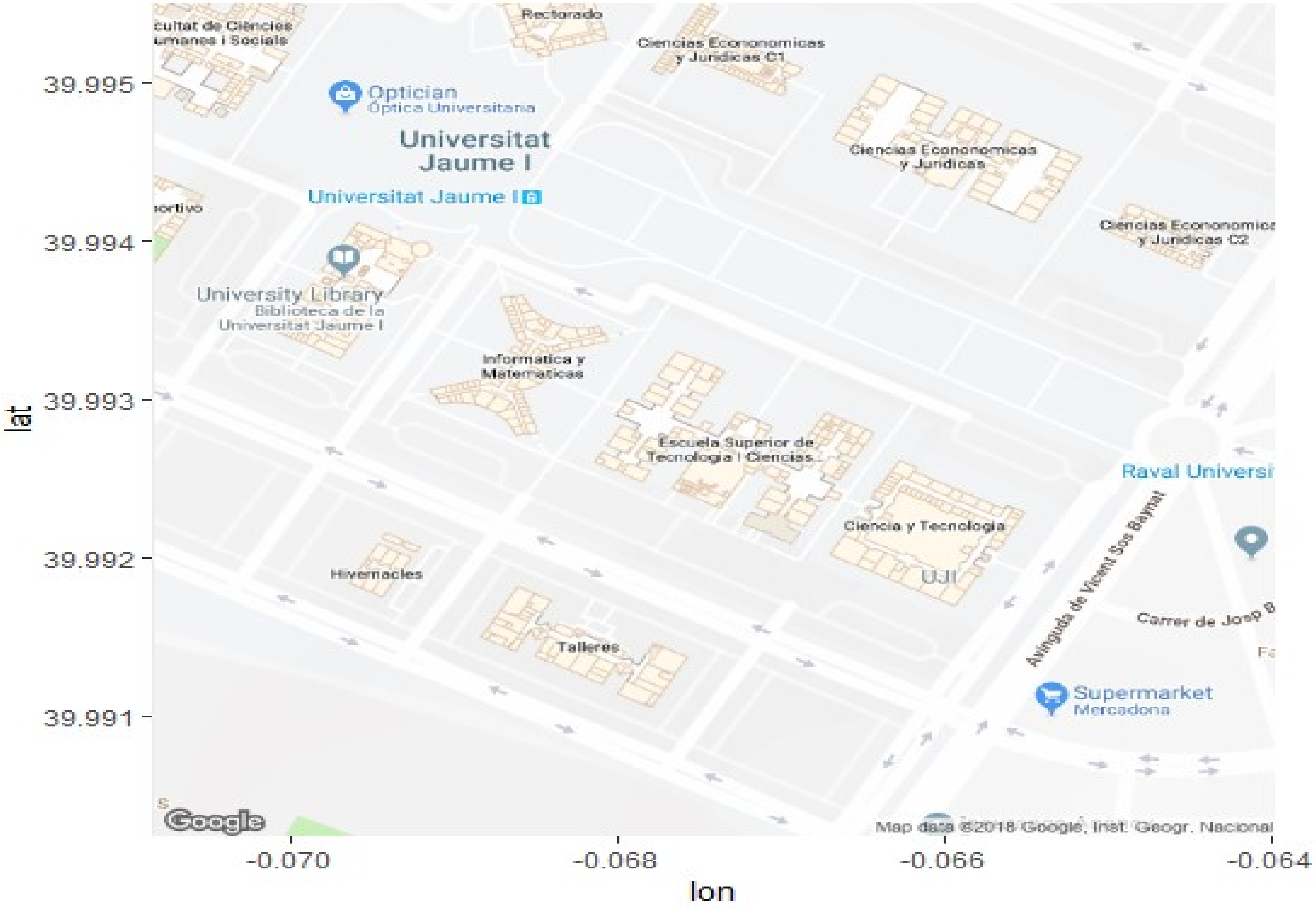
DATA ANALYTICS OBJECTIVES

- Develop a Wi-Fi Fingerprinting model to locate people Indoors in 3 large buildings in a campus, Evaluate the data analytics models by predicting a person's location
- Use Wi-Fi fingerprint characterized by Wireless Access Points (WAPs) and the corresponding RSSI: Received Signal Strength Intensity
- Compare of the models produced by at least three different algorithms
- Recommend the algorithm that is best suited for this data and a justification why it is the preferred choice
- Recommend improvements that can be achieved, based on research on indoor locationing or experimentation with the dataset.

Indoor Location Dataset

- Automatic user localization consists of estimating the position of a user by using an electronic device, usually a mobile phone
- Our Dataset is focused on WLAN fingerprint-based technologies and methodologies, known as Wi-Fi Fingerprinting
- UJIIndoorLoc Dataset is presented a common database for comparison of models produced by different algorithms
- This Indoor Location Dataset covers three buildings of Universitat Jaume I with 4 or more floors
- The dataset has 529 attributes

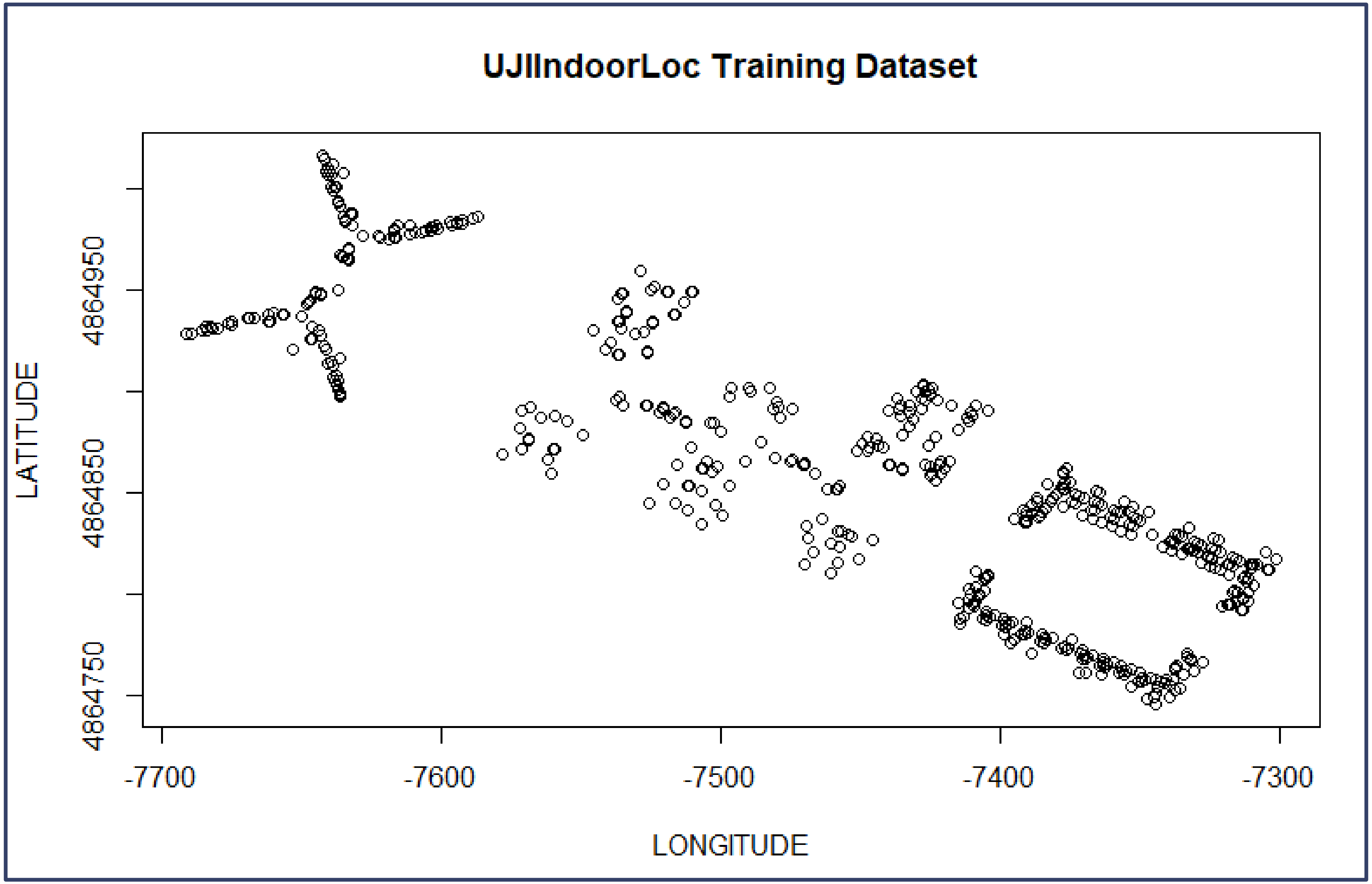
Universitat of Jaume I in real life



Description of the Indoor Location Dataset

- Attribute 001 (WAP001) through Attribute 520 (WAP520): Intensity value for WAP001 through WAP520. Negative integer values from -104 to 0. The value +100 is used when WAP001 was not detected.
- Attribute 521, LONGITUDE : Negative real values from -7695.93875493 to -7299.78651673
- Attribute 522, LATITUDE : Positive real values from 4864745.7450159714 to 4865017.3646842018.
- Attribute 523, FLOOR : Floors inside the building. Integer values from 0 to 4.
- Attribute 524, BUILDINGID : ID to identify the building. Integer values from 0 to 2.
- Attribute 525, SPACEID : Internal ID number to identify the Space, indicating office, corridor or classroom
- Attribute 526, RELATIVEPOSITION : Relative position with respect to the Space, 1- Inside, 2- Outside.
- Attribute 527, USERID : User identifier. Categorical 3-digit integer values, random between 158 and 186
- Attribute 528, PHONEID : Android device identifier Categorical integer values.
- Attribute 529, TIMESTAMP : UNIX Time when the capture was taken. Integer value.

UJIIndoorLoc Training Dataset



Evaluation of Data

- Defined LOCATION attribute that consists of BUILDINGID, FLOOR, SPACEID
- TRAINING Dataset was split into Training and Validation as 70% & 30%, respectively
- Separated data by BUILDINGID for Training and Validation
- Selected kNN, Random Forest and C5.0 as the algorithms to build models
- Each algorithm was evaluated using 10-fold cross validation
- WAP attributes from WAP001 to WAP520 are used as predictors, where each WAP corresponds to a RSSI Received Signal Strength Intensity.
- The intensity values are represented as negative integer values ranging -104dBm (extremely poor signal) to 0dbM.
- RSSI +100 is used to denote when a WAP was not detected

Preprocess Data

- Dataset had 19937 observations of 529 variables
- Checked for NA values, none found
- Removed duplicate records to reduce dataset to 19300 records
- BUILDINGID, FLOOR, SPACEID are defined as integer
- LONGITUDE, LATITUDE are defined as numeric
- TIMESTAMP is defined as POSIXct in UTC time zone
- LOCATION is the feature used in classification solution ---- it is converted to factor

Training and Testing Process

- Split TRAINING Dataset into Training and Validation (Testing) set as 70% & 30%, respectively
- Further reduced the Training data by filtering down to 2000 random records as the subset
- Separated data by BUILDINGID for Training and Validation
- Selected kNN, Random Forest and C5.0 as the algorithms to build models
- Each algorithm was evaluated using 10-fold cross validation
- WAP attributes from WAP001 to WAP520 are used as predictors, where each WAP corresponds to a RSSI Received Signal Strength Intensity
- RSSI values are represented as negative integer values ranging -104dBm to 0dbM. RSSI +100 is used to denote when a WAP was not detected
- LOCATION is the dependent variable that we predict.

Algorithms to evaluate 3 models

To get the best parameters, each algorithm was tested with different parameter options in multiple R script runs:

1. kNN was evaluated using
   i. k = 5, 7, 9, 11, 13, 15, 17 *(first run)*
   ii. k = 1, 2, 3, 4 *(final run)*

   The final value used for the model was k = 1 using Accuracy & Kappa to select the optimal model.

2. Random Forest was evaluated using
   i. mtry = 1, 2, 3, 5 *(first run)*
   ii. mtry = 16, 32, 48 *(final run)*

   The final value used for the model was mtry = 32 using Accuracy & Kappa to select the optimal model.

3. C5.0 was evaluated using Boosting Iterations (trials)
   i. trials = 2, 8, 24 *(first run)*
   ii. trials = 32, 48, 64 *(next run)*
   iii. trials = 48, 96 *(final run)*

   The final values used for the model were trials = 96, model = rules and winnow = FALSE, using Accuracy & Kappa to select the optimal model.

METRICS for the RESULTS

- Accuracy and Kappa are gathered as the two performance metrics from each model

ACCURACY is calculated as the proportion that represents the number of true positives and true negatives, divided by the total number of predictions

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN)$$

		Predictions	
		no	yes
Actuals	no	TN	FP
	yes	FN	TP

- **KAPPA** adjusts ACCURACY by accounting for the possibility of correct prediction by chance alone. A common interpretation is shown as follows*:

- The criteria for the "best model" is the highest value

  for Accuracy and Kappa when predicting the LOCATION

Kappa:			
Poor agreement	< 0.20		
Fair agreement	0.20	to	0.40
Moderate agreement	0.40	to	0.60
Good agreement	0.60	to	0.80
Very Good agreement	0.80	to	1.00

* Machine Learning with R, Brett Lantz

- Accuracy and Kappa are obtained by post resampling predicted values against actuals in the validation/testing subset, which is 30% of the original Training dataset in this project.

RESULTS:

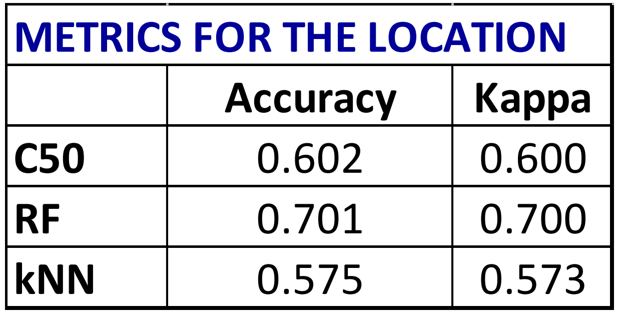
MODEL ALGORITHMS, PARAMETERS and METRICS of PREDICTING LOCATION

Algorithms	Run	Model Parameter (tunegrid)	Parameter Values	Parameter Value Selected Model	Accuracy	Kappa
kNN	Last	k values	k = 1, 2, 3, 4	k = 1	0.575	0.573
RF	Last	mtry	mtry = 16, 32, 48	mtry = 32	0.701	0.700
C50	Last	trials	trials = 48, 96	trials = 96	0.602	0.600
C50	Next	winnow, trials (Boosting Iterations)	winnow = (TRUE, FALSE), trials=(32, 48, 64), model=("tree", "rules")	winnow = FALSE, trials=24, model="rules"	0.601	0.599
kNN	First	k values	k = 5, 7, 9, 11, 13, 15, 17	k = 5	0.465	0.462
RF	First	mtry	mtry = 1, 2, 3, 5	mtry = 5	0.587	0.585
C50	First	winnow, trials (Boosting Iterations)	winnow = (TRUE, FALSE), trials=c(2, 8, 24),	winnow = FALSE, trials=24	0.587	0.584

RESULTS:

- Random Forest (mtry = 32) is selected as the best model to predict LOCATION, and is recommended as the best algorithm to determine a person's location in indoor spaces using UJIIndoorLoc Dataset, based on the Accuracy and Kappa values after the final run of the algorithms

The Accuracy and Kappa for LOCATION are calculated using the Weighted Average of the Accuracy and Kappa for each Building, using Number of Levels as the Weight.

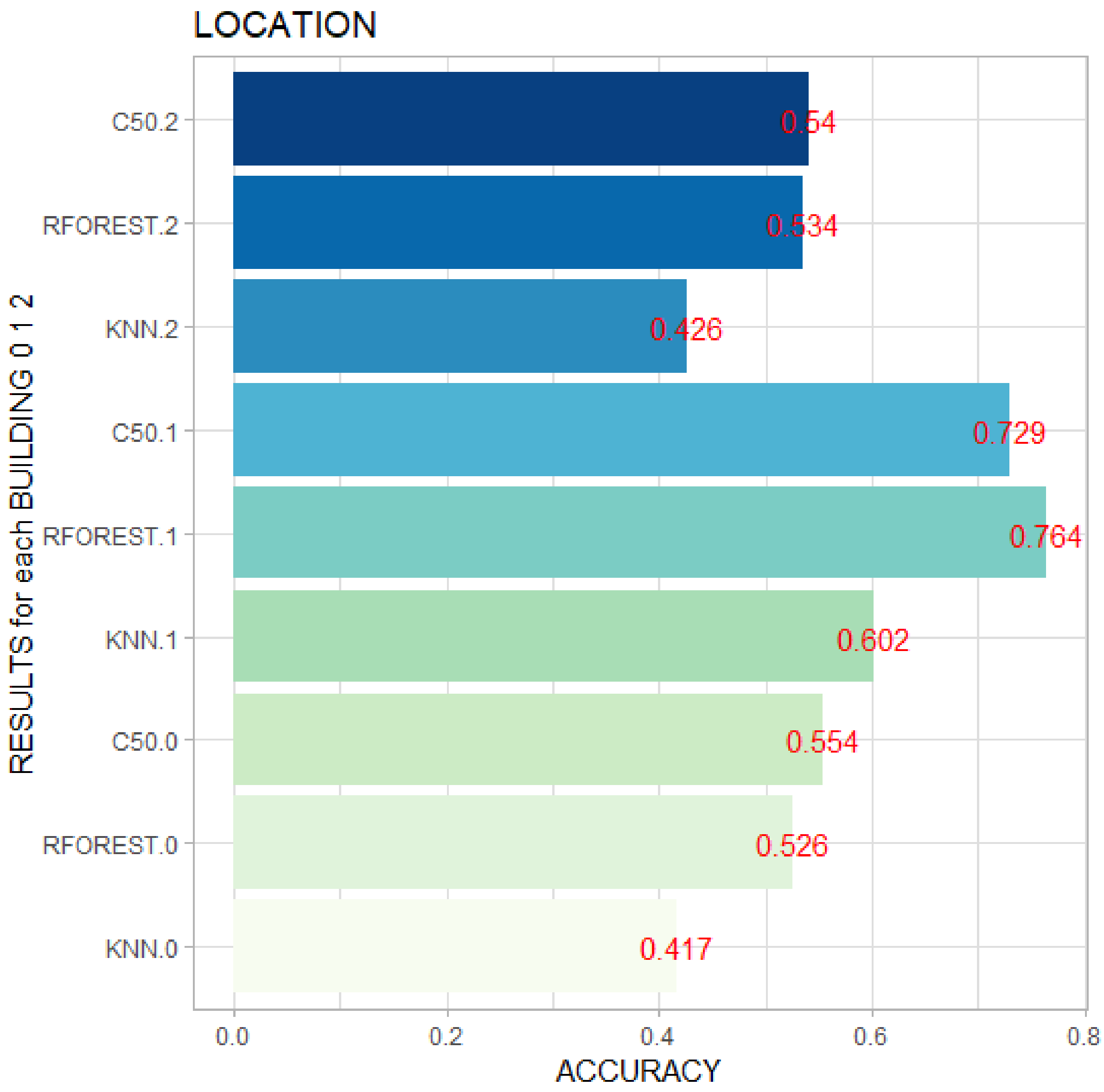


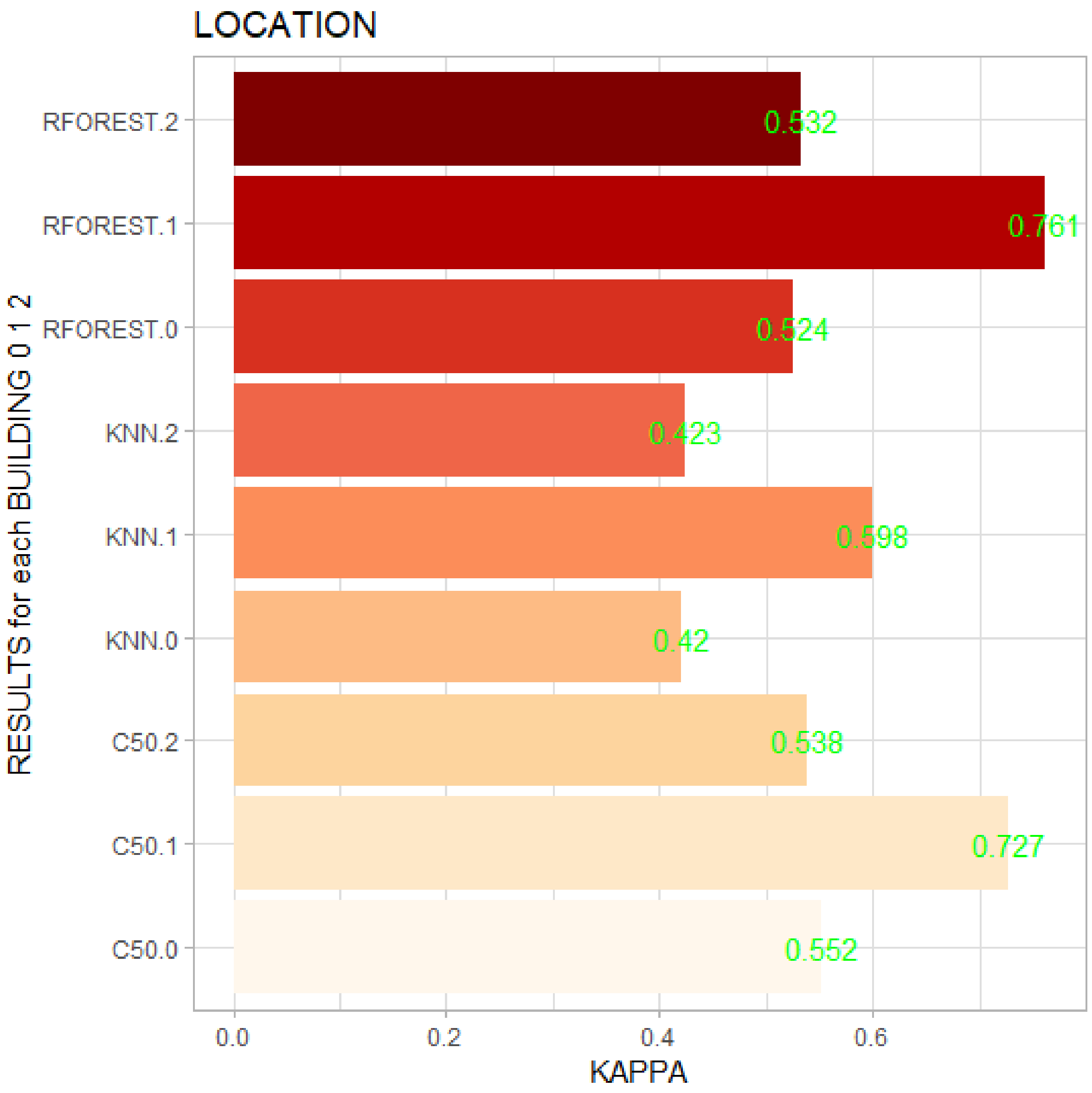
METRICS FOR THE LOCATION

	Accuracy	Kappa
C50	0.602	0.600
RF	0.701	0.700
kNN	0.575	0.573

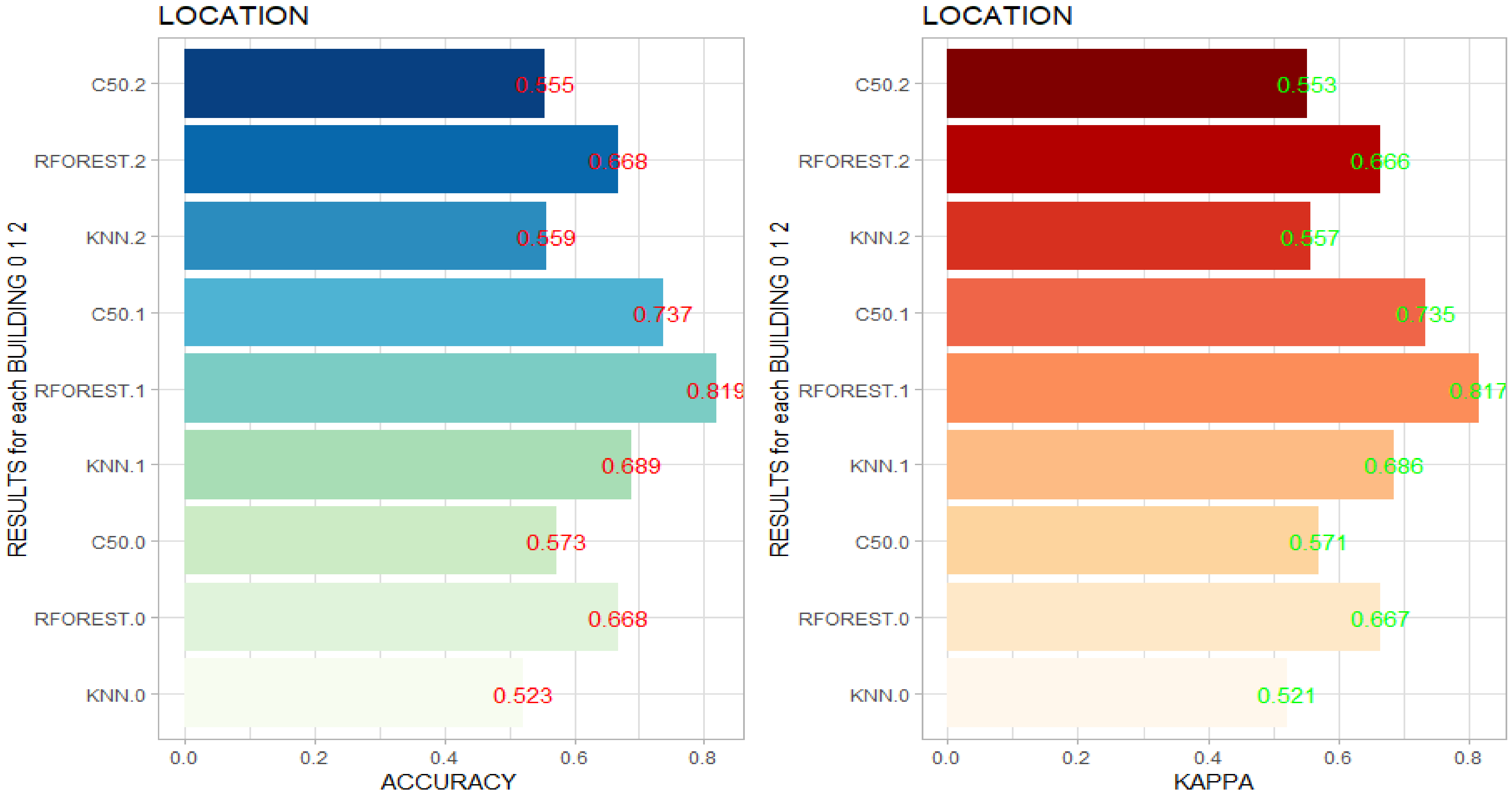
- Details of model results using each algorithm are provided in an accompanying EXCEL spreadsheet with the same *filename*.xlsx as this PowerPoint file.
- The Excel file shows results from each algorithm after the first run, and after the final run.
- Plots following this slide shows:
  - Accuracy and Kappa for each algorithm (kNN, RF, C5.0) for the prediction of Location by Building after the first run and the final run.
  - Accuracy and Kappa for each algorithm for the prediction of LOCATION (BuildingID, Floor and SpaceID).

Metrics for each Model on the Prediction of LOCATION by Building (1^{st} run)

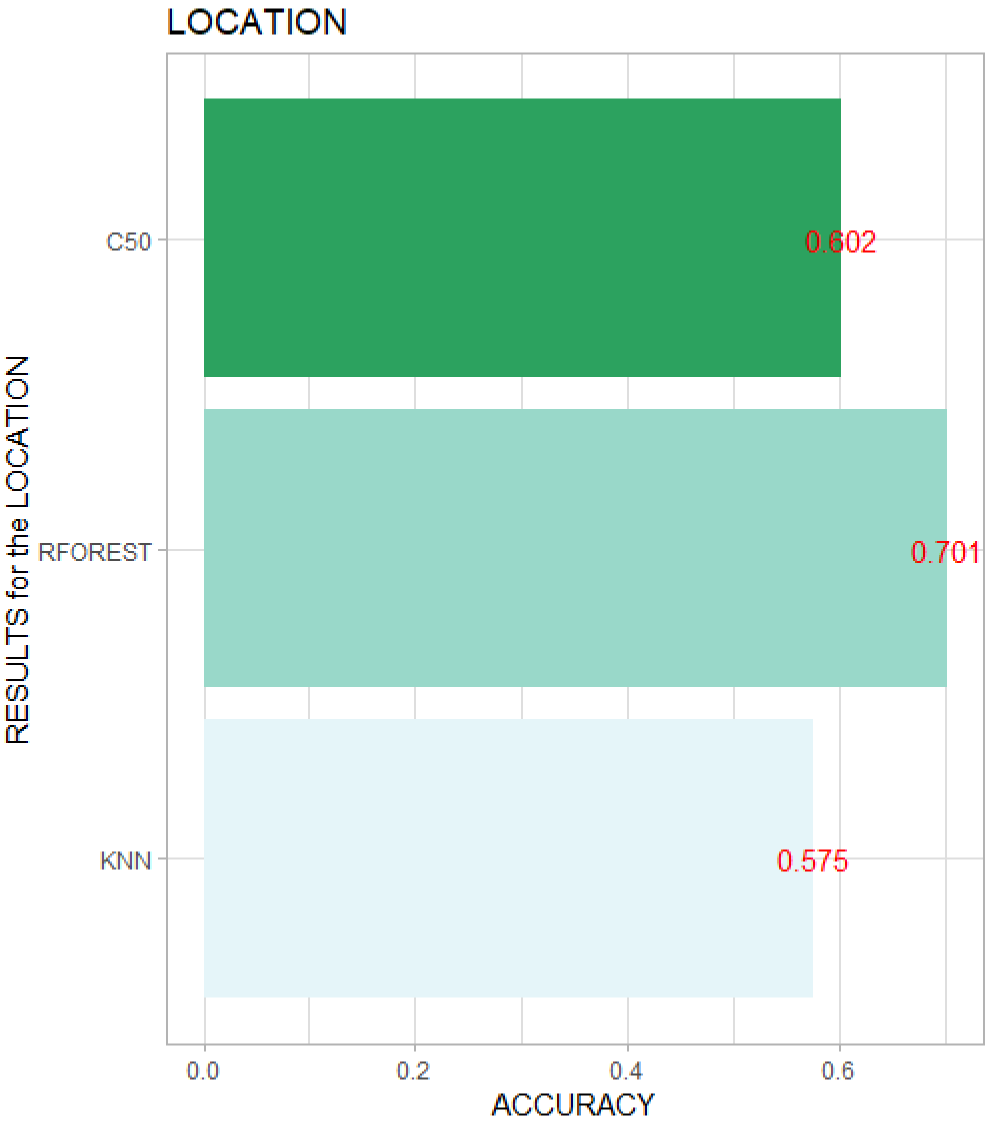


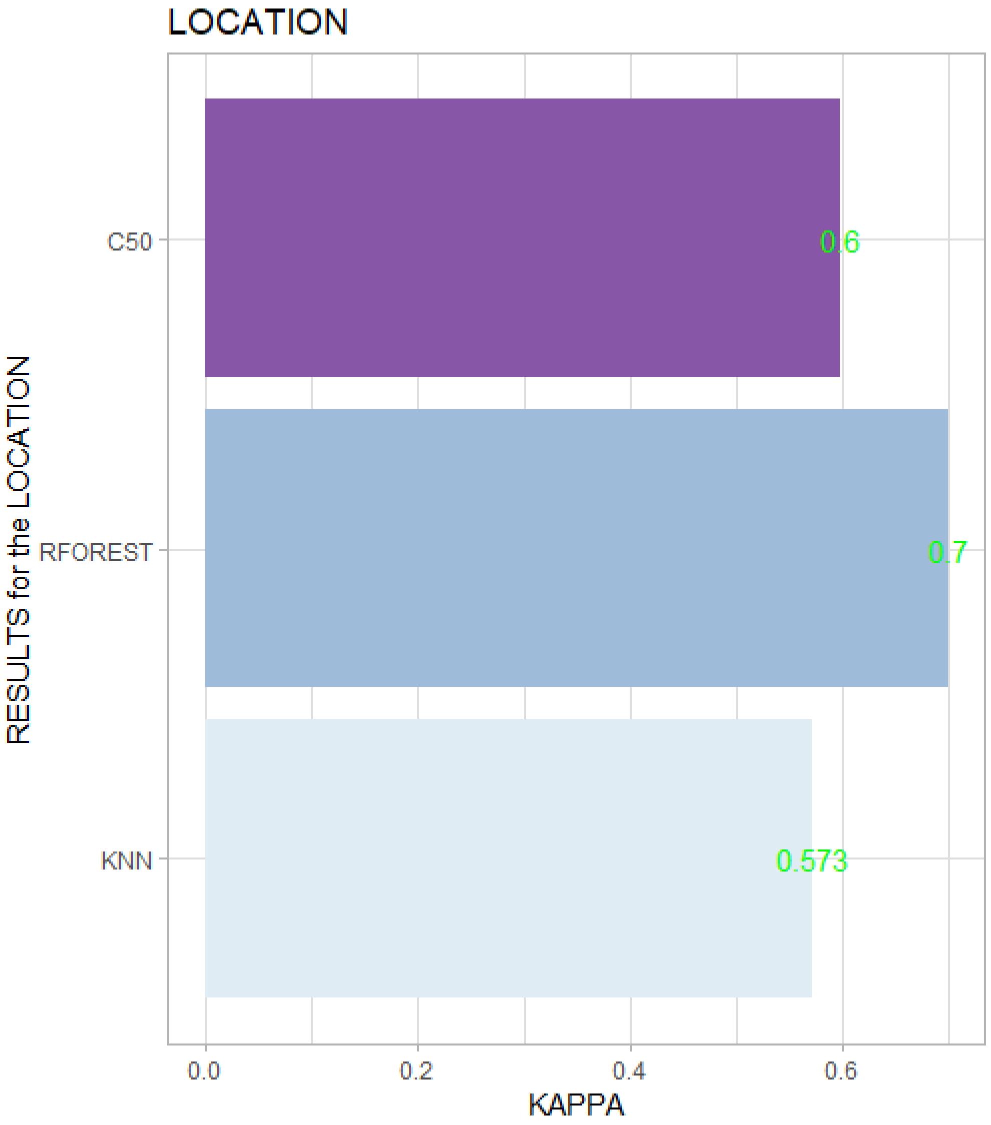


Metrics for each Model on the Prediction of LOCATION by Building (Final run)



Accuracy & KAPPA for Prediction of LOCATION (Final Results)





RECOMMENDATIONS

1) Developed a useful model that would help people navigate large, complex, unfamiliar interior spaces. We showed how this model worked with the UJI Campus dataset. The campus is 110 000 m^2. We showed that our Wi-Fi fingerprinting model can locate a person in an office, classroom or corridor – space with a typical area= ~400 m^2. Our results indicate that in UJI Campus, we can locate 7 out of 10 people (based on LOCATION Accuracy=0.701) and pinpoint their location, the size of which is only 0.36% of the total areal extent of the search area. Recommendation is to promote this Wi-Fi locationing system since many businesses depend on where people gather up and location-based marketing.

2) The dataset was created using Android devices. Different devices with other mobile operating systems such as Apple iOS, Blackberry OS, MS Windows Mobile should be included in an updated database.

3) The model and results could be improved by testing it on another dataset that includes amount of activity (# of visitors) and resulting profit in locations such as hotels or restaurants in pre-selected areas of a city.

4) Data acquisition for the indoor positioning systems must include permission of the person who owns the device. For example, Find Friends app that uses GPS for outdoor localization requires explicit permission before one can use the app to locate another person.

5) Results and solution procedure could be improved by reducing the dimension of the problem, for example, predicting LOCATION by taking into account the WAP with the highest RSSI.