Technical Report

Evaluate Techniques for Wifi Locationing

XTOL Data Analytics and Big Data program

Module 3, Task 3

Esteban Villalobos Gomez

November 14_{nd}, 2019

Contents

Goals

- The main goal is to analyze and predict a user location inside the Universitat Jaume I, based on Wireless Access Points (WAP) readings.
- Train and compare at least three different models to predict a user's location, defined as Building number and Floor number.

Description and location of related data sources

- The training and validation datasets are located at https://archive.ics.uci.edu/ml/datasets/UJIIndoorLoc.

    - The dataset contains 19937 training observations (trainingData.csv file) and 1111 validation records (validationData.csv file).
    - The dataset covers three buildings of Universitat Jaume I with at least 4 floors.
- It consist of 529 variables, from which 520 are different WAPs, the other nine are (taken description from link:

  Attribute 521 (Longitude): Longitude. Negative real values from -7695.9387549299299000 to -7299.786516730871000

  Attribute 522 (Latitude): Latitude. Positive real values from 4864745.7450159714 to 4865017.3646842018.

  Attribute 523 (Floor): Altitude in floors inside the building. Integer values from 0 to 4.

  Attribute 524 (BuildingID): ID to identify the building. Measures were taken in three different buildings. Categorical integer values from 0 to 2.

  Attribute 525 (SpaceID): Internal ID number to identify the Space (office, corridor, classroom) where the capture was taken. Categorical integer values.

  Attribute 526 (RelativePosition): Relative position with respect to the Space (1 - Inside, 2 - Outside in Front of the door). Categorical integer values.

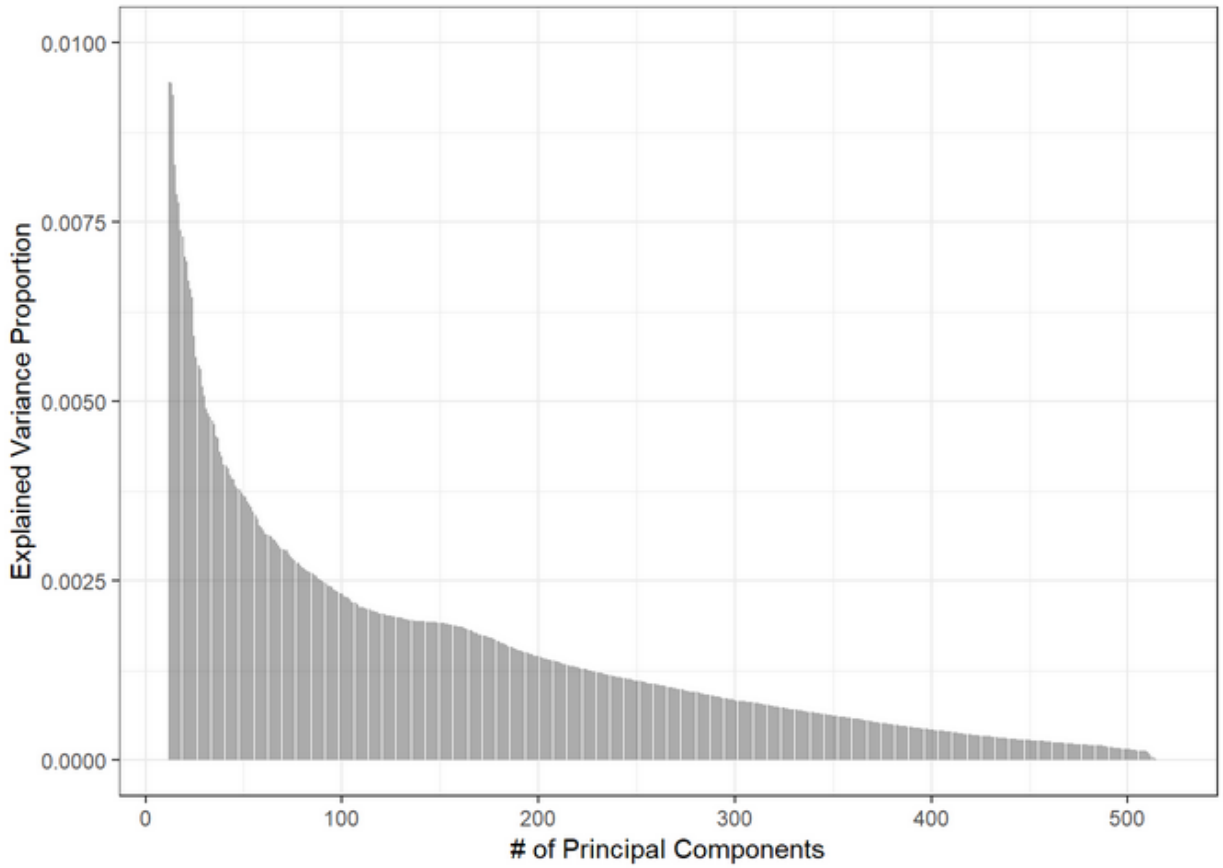
  Attribute 527 (UserID): User identifier (see below). Categorical integer values.

  Attribute 528 (PhoneID): Android device identifier (see below). Categorical integer values.

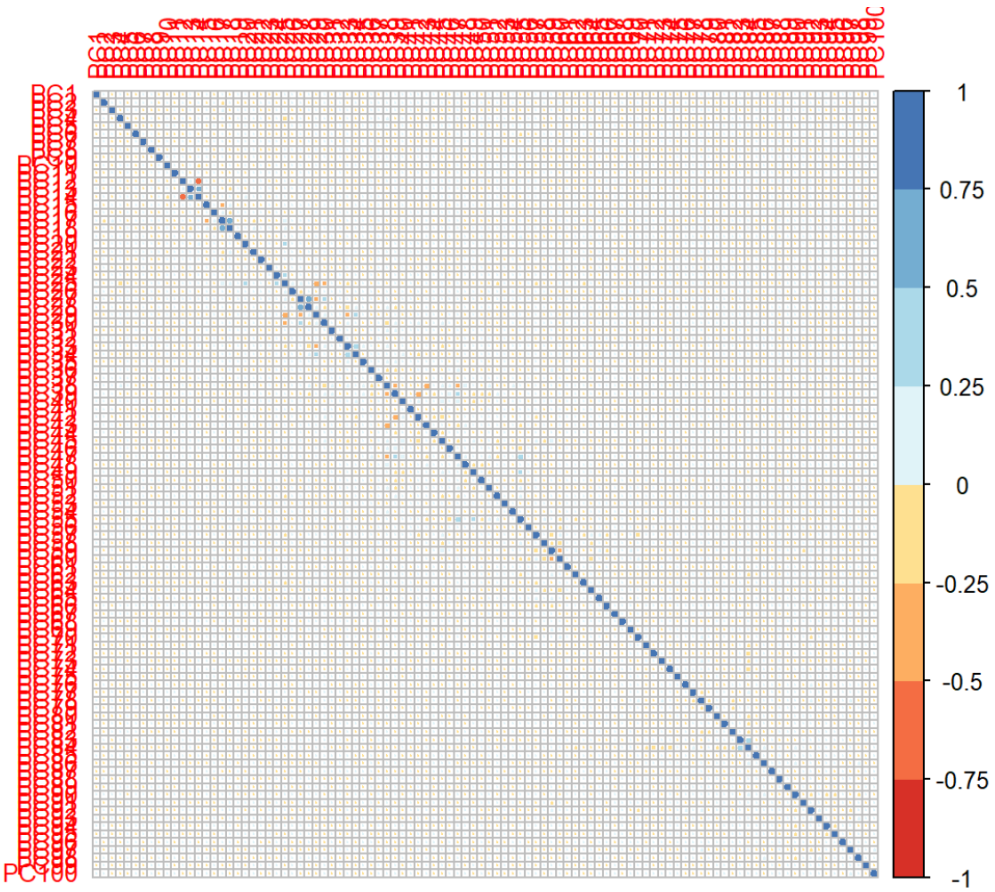
  Attribute 529 (Timestamp): UNIX Time when the capture was taken. Integer value.

- In order to train the models, the WAP readings will be treated as the **independent** variables and the BuildingID and Floor ID as the **dependent** variables.

- Given that 520 variables (features) will cause the models to take a long time to train, the **Principal Component Analysis** (PCA: https://rpubs.com/Joaquin_AR/287787) was used to reduce dimensionality to just the top 100 principal components that would explain 60% of the variability of the dataset as shown below:



- All the top 100 Principal components show low correlation between each other:



- Two sets of models were trained, the first set only predicted the BuildingID, the second set predicted both Building and Floor at the same time.
    - The second set of models took longer to train, and they were the ones I selected for this project.
    - Once the models were trained, they were saved into disk, and logic was implemented to load them again for faster notebook regeneration.

Important Note: The source code for training and validating the models is attached in the R notebook: **wifi_fingerprintin_v2.Rmd**

How data will be managed for the project

Incoming data will be managed in AWS, stored in a NoSQL database and when requiring a prediction, the readings will be injected into a a streaming framework like AWS Kinesis, and will be used as the input for the selected classifier.

Known issues with the data and your planned solutions

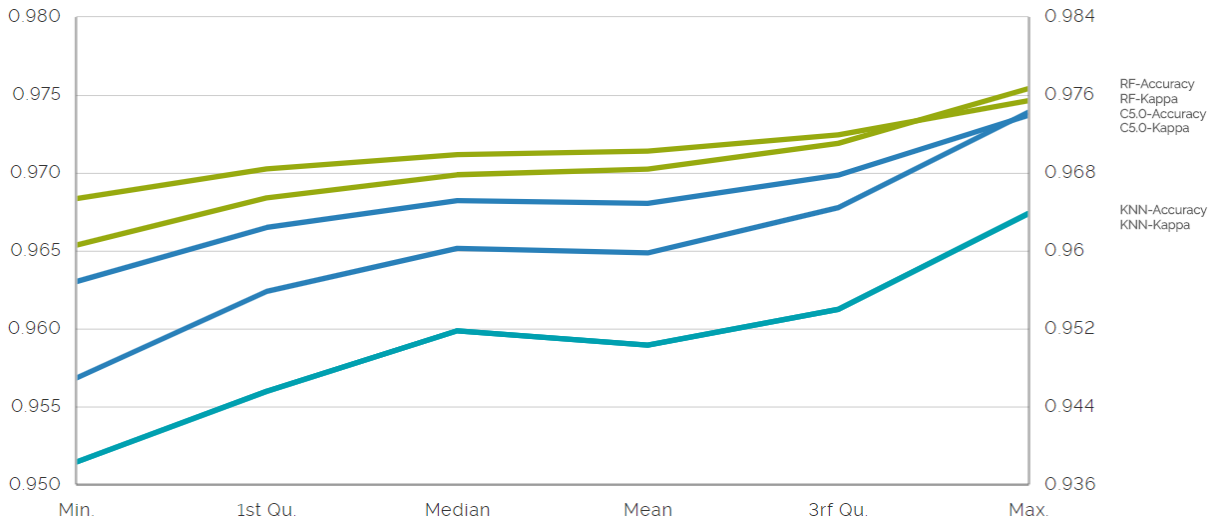
- The validation set is missing several information, like the SpaceID, which prevented me to train and validate a classifier to predict the specific section as well as the floor and the building.
- The training set is biased towards the areas students most frequently visit, and is missing information on many sections of the different buildings, like laboratories the students do not have access to or other restricted areas.
    - Many WAP didn't record any signals because of this issue.

Comparison of the models produced by at least three different algorithms

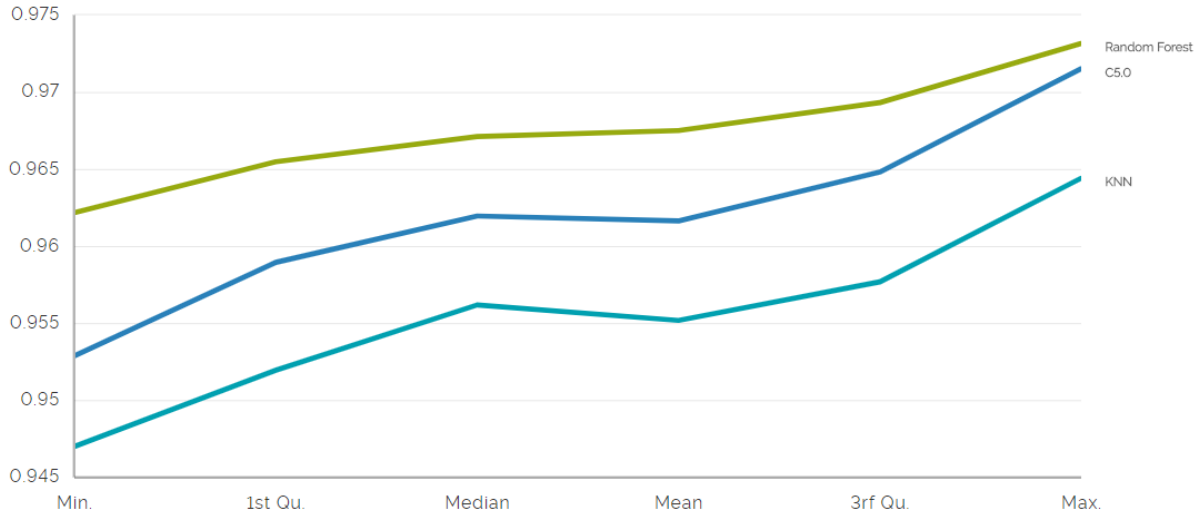
The models trained models predict both the Building and Floor (second set of models in the source code).

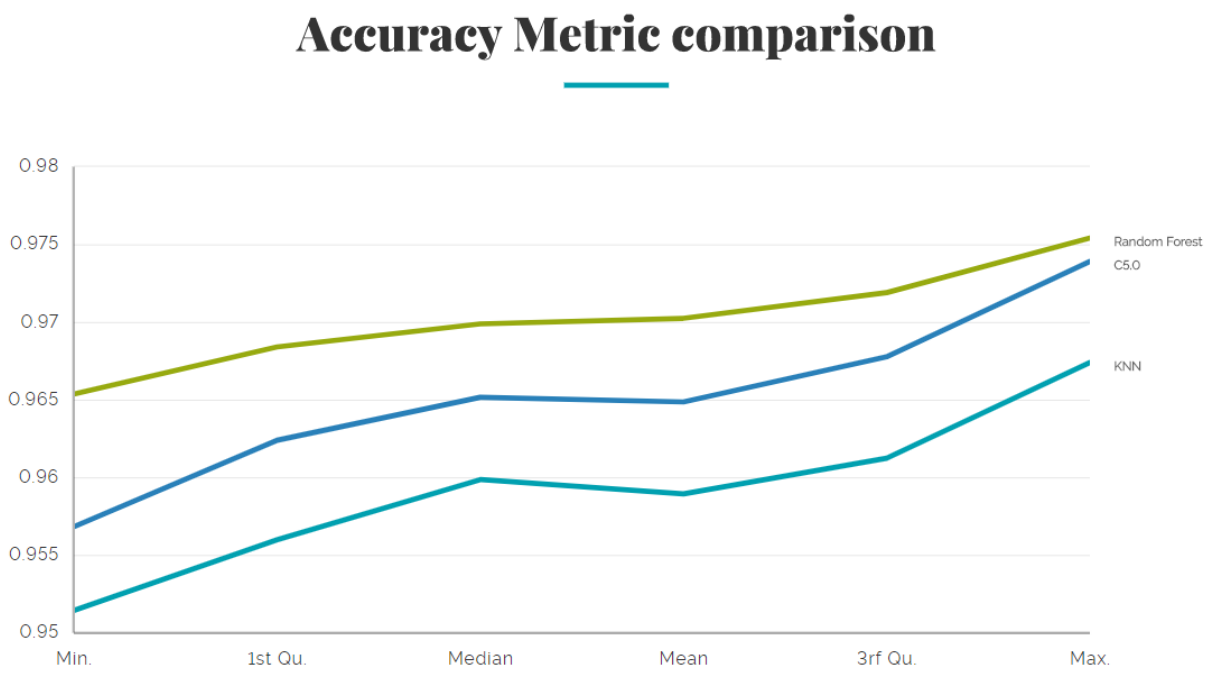
The following table shows both the accuracy and kappa metrics for the three trained models I selected, which were K-Nearest-Neighbors, Random Forest, and the C5.0 algorithm:

Kappa and Accuracy comparison



Kappa Metric comparison





Algorithm recommendation

By evaluating the accuracy and mainly the Kappa metric (because is usually less misleading than accuracy) the recommended model to use for classification is the **Random Forest**, this is because both metrics performed better than the other two algorithms.

Recommendations on indoor locationing

In order to improve predictions:

1. Increase the training and validation set with more SectionIDs, so that this variable can be predicted as well.
    a. Also ensure more data is taken from most of the locations within the University buildings.
2. When predicting validate the several predicted locations with a timestamp, so that we can discard predictions that wouldn't be possible (like predicting a user's location in two different buildings within seconds from each reading).
3. Combine the WAP signals with other signals in order to create a Hybrid positioning system:
    a. Other input signals could be taken from: cell tower signals, Bluetooth sensors, IP addresses and network environment data.

b. NFC could be used as well to fingerprint users, sensors could be collocated at some "choke points", like doors or elevators.
c. More advanced techniques could be implemented to give a very accurate prediction, like the **Angle of arrival**: "AoA is usually determined by measuring the time difference of arrival (TDOA) between multiple antennas in a sensor array. In other receivers, it is determined by an array of highly directional sensors—the angle can be determined by which sensor received the signal. AoA is usually used with triangulation and a known base line to find the location relative to two anchor transmitters.". Source: https://en.wikipedia.org/wiki/Indoor_positioning_system#Magnetic_positioning

4. Another option is to use a currently available locationing solution like:
   a. Google Indoor Maps: https://www.google.com/maps/about/partners/indoormaps/
   b. Insoft's Indor Navigation: https://www.infsoft.com/solutions/indoor-navigation