SEMANTIC PLACE PREDICTION FROM MOBILE PHONE DATA

(AKILLI TELEFON VERİSİ KULLANARAK ANLAMSAL YER BELİRLEME)

by

Selek Ceren ÇELİK, B.S.

Thesis

Submitted in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

in

COMPUTER ENGINEERING

in the

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

of

GALATASARAY UNIVERSITY

June 2016

This is to certify that the thesis entitled

SEMANTIC PLACE PREDICTION FROM MOBILE PHONE DATA

prepared by **Selek Ceren ÇELİK** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering** at the **Galatasaray University** is approved by the

Examining Committee:

Assist. Prof. Dr. Özlem DURMAZ İNCEL (Supervisor)
Department of Computer Engineering
Galatasaray University ------------------------

Prof. Dr. Şebnem BAYDERE
Department of Computer Engineering
Yeditepe University ------------------------

Assist. Prof. Dr. Ahmet Teoman NASKALİ
Department of Computer Engineering
Galatasaray University ------------------------

Date: ------------------------

ACKNOWLEDGMENTS

I want to express my gratitude to my advisor Ozlem Durmaz Incel for guiding me through the thesis process with her knowledge and understanding . Also, I wish to thank my family and friends for supporting me with their kindness and patience. This research is supported by Tubitak under the project number 113E271 and by the Galatasaray University Scientific Research Projects (BAP) under the grant agreement number 15.401.004

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Semantic place prediction problem is the process of giving semantic names, such as school, home, office, to locations. Different than the localization problem where the coordinates of a place are predicted, the aim is to semantically characterize the location. While the GPS coordinates of a place can be utilized in solving the problem, phone usage patterns of the users in that location can be used as well. In this thesis, we collected data from 20 participants from Galatasaray University for a month-duration. Our aim is to semantically classify locations of smart phone users utilizing the data collected from wireless interfaces and the motion sensors available on the phones with machine learning algorithms. The efficiency of features extracted from raw data is analysed in terms of metrics such as accuracy, using different classification algorithms such as Decision Tree, Random Forest, K-Nearest Neighbour and Naive Bayes. Using attributes selection algorithms we discovered the features from communication, time and activity feature groups returned the highest success rates. The results were almost 80% for cross validation and close to 100% in terms of accuracy with leave one subject out using Random Forest classification algorithms.

Keywords : semantic, place prediction,mobile phone

ÖZET

Semantik yer tahmin problemi yerlere, okul, ev, ofis olarak semantik isimler verme sürecidir. Yerelleştirme sorunu olan bir yerin koordinatları tahmin edilmesinden farklıdır, amaç anlamsal olarak konumu karakterize etmektir. Bir yerin GPS koordinatları sorunun çözümünde kullanılabilir olsa da, istenen konumdaki kullanıcıların telefon kullanım alışkanlıkları da kulanılmış olur. Bu tezde, bir ay süreyle Galatasaray Üniversitesi'nden 20 katılımcı veri toplanmıştır. Amacımız kablosuz arayüzleri ve makine öğrenme algoritmaları ile telefonlardaki kullanılabilir hareket sensörleri ile toplanan verileri kullanılarak akıllı telefon kullanıcılarının konumlarını anlamsal olarak sınıflandırmaktır. Ham verilerden çıkarılan özelliklerin verimliliği Karar Ağacı, Random Forest, K-En Yakın Komşu ve Naif Bayes gibi farklı sınıflandırma algoritmaları kullanarak, doğruluk gibi metrikler açısından analiz edilir. Öznitelik seçim algoritmaları kullanarak iletişim, zaman ve aktivite öznitelik gruplarından yüksek başarı oranları elde edilmiştir. Sonuçlarda Rastgele Orman sınıflandırma algoritmasını kullanarak çapraz doğrulama ve bir-kişiyi-dışarıda-bırak (BKDB) methodları ile, sırasıyla, 80% ve 100% başarı elde edildi.

Anahtar Kelimeler : anlamsal, yer tahmini, mobil telefon

1 INTRODUCTION

Context of a user can involve various information both about the user and his/her surrounding, including user's activities, his/her location, emotions, etc. Context-aware applications adapt their behaviour according to the context of the user. For example, if the person is to be detected in a meeting, the incoming calls can be declined, or if the person is detected in a noisy environment, volume of the phone can be increased. In order to provide context-aware services first the context of the user should be predicted and in this regard, the mobile devices, such as the phones and smart watches, provide a unique platform for context recognition applications with the integrated rich set of sensors, such as GPS, accelerometer, their ubiquity, ease of use and wireless communication capabilities with various interfaces.

Location is one of the key aspects of a user's context. If the location of a user is known, suggestions based on the location information can be presented, such as the closest branch of a bank. Location-based services (LBS)use location data to control services presented to a user. However, in terms of privacy, revealing a user's physical location is a sensitive issue Hence, privacy-sensitive solutions can be used to tackle with this problem in LBS.

Semantic place prediction problem is the process of giving semantic names, such as school, home, office, to locations. Different than the localization problem where the coordinates of a place are predicted, the aim is to semantically characterize the location. While the GPS coordinates of a place can be utilized in solving the problem, as mentioned due to privacy reasons, some users may not want to share their exact location information and disable location settings. On the other hand, phone usage patterns of the users and their performed activities in that location can also be very useful in semantically characterizing the places. For instance, in an office environment a user may be mostly still, however when he is transporting he will be mobile.

In this thesis, the aim is to semantically classify locations of smart phone users utilizing the data collected from wireless interfaces, the motion sensors available on the phones and phone usage characteristics (number of running applications, battery level, etc.) with supervised machine learning algorithms. For this purpose, we collected data from

17 different participants, with a minimum duration of 1 month. In the data collection phase, we utilized an application named ARService (Coskun, 2014), which logs sensor data from smart phones. The application collects data from the integrated sensors on mobile phones, from the usage of wireless interfaces (WiFi and cellular data), as well as phone usage, such as running applications, battery level, screen on/off transitions. Additionally, the application periodically predicts the activity of the user by utilizing data from accelerometer. The detected activities are sitting, standing, walking, running, transportation and stairs. The collected data is uploaded to a server for further analysis. In order to be able to train our classifiers, the application also collects place label (class) information from the users. The users can mark up their location, similar to Foursquare check-ins. Additionally, the application asks the user about his location periodically, every 15 minutes.

After collecting the data with the ARService application, next we pre-processed the data for combining different sets of tables and extracted features. For each user, ARService creates 4 tables, each containing data from different sources : communication information, activity information, phone usage information and place labels. These sets should be merged to identify features. By using the timing information available in each table, we could merge the data. As the next step, we extracted features from the raw data. We identified the data which could be important for place prediction and categorized the features into three classes. The first class includes features related to time. It is clear that people have daily routines, such as being at home at night, and timing information is a key aspect in predicting a place. Related to timing information we extracted features including day of the week information, week/weekend information, time of the day information and period of day (morning, afternoon, etc.).

The second class of features is extracted from communication information. As mentioned, ARService logs data from wireless interfaces. It scans the nearby WiFi access points and collects data from the cellular network about the nearby base station id's. The number of access points and base stations change according to the place of the user and they can be good predictors in identifying the places. Instead of using exact access point and base station ids, we used the number of unique id's nearby a user. Using exact id information would be user specific and hence would limit the generalization of our methodology to new users.

As the third class of features, we utilized the activity data. People perform specific activities in specific locations. While sitting and standing can be the dominant activities in an office environment, running or walking could be the dominating activities while outdoors or in a gym. ARService predicts the activities of the users every seconds and as mentioned the list of activities include six different motion activities. We used the ratio of each activity in a specific location.

As the final class, we used the phone usage data. This data includes the number of running applications, battery level, whether the headset is on/off, and whether the screen is on/off. Phone usage patterns also change according to the location. While a user may not interact with the phone in an office environment, he/she may have more time to use the phone at home or in a cafe.

All these feature sets are combined into a single table by merging information from different sources of data collected from ARService. The final step of our methodology includes the classification/prediction step. For this purpose, we utilized supervised machine learning algorithms, including decision tree, random forest and KNN. in Weka machine learning toolkit (Hall et al., 2009). These classifiers have been used in the literature (Do and Gatica-Perez, 2014; Zhu et al., 2013) and are reported to perform well, so that we can compare our performance with related studies. The list of labels is as follows :

1. Home,
2. Friend's/parent's home,
3. Work/School,
4. On the Road (Transportation),
5. Outdoor (Park, etc.),
6. Canteen/Restaurant/Cafe/Bar,
7. Mall/Shop,
8. Other,
9. GSU Classroom/Lab,
10. GSU Canteen/garden,
11. GSU Library.

This list is formed according to the feedback we gathered from our users about their significant places. Additionally, the users can add any location which is not given in the list by the ARService application.

In the initial set of experiments we collected data from 3 individuals in order to analyse the efficiency of features extracted from raw data in terms of metrics such as accuracy, using different classification algorithms. The results show that, while random forest and decision tree algorithms achieve 66% accuracy with only time features, adding features from communication features and activity features increases the accuracy up to 99%. In the second round, we collected data from 14 participants with an age range of 18 to 40, including Galatasaray University students and faculty members. Similar to the initial study, we analysed the performance of feature sets with different classifiers. After analysing the features we focused on the impact of validation methodology in the classification phase. In the initial analysis, cross-validation per user is reported and next its performance is compared with cross validation for all users and leave-one-subject-out (LOSO) methods. Although, the place characteristics may differ according to the users, the LOSO method enables us to evaluate the generalization of our solution to new users without any training data.

The following lists the main contributions and highlights of this thesis.

- We have collected mobile data from more than 20 participants with a duration of month.
- We focus on the efficiency of features in predicting places and show that while time features may not perform well when used alone, when they are combined with features extracted from user's activities, communication and phone usage patterns, the accuracy of prediction increases.
- We visualize the data and show the dominant features for place prediction.
- Different than previous studies, we also show that activities performed in specific places give important clues in predicting these places.
- Many of the semantic place prediction works collect the GPS data, social network markings, etc. to predict next place. Which causes to privacy violation. In our research we try to prove, that one can make a high rate prediction without disclosing location information.

The remainder of the thesis is organized is as follows : In Chapter 2 we review the existing studies from the literature and then explain the tools that are used in the implementation. Chapter 3 explains the details of our methodology, including data collection, feature extraction and classification. In Chapter 4, we evaluate the performance of our methodology according to different metrics, including features and validation methods. Chapter 5 concludes the thesis and explains the future work for further studies.

2 LITERATURE REVIEW

In this section, first we review the existing studies from the literature and then explain the tools that are used in the implementation.

2.1 Related Work on Mobile Data Collection

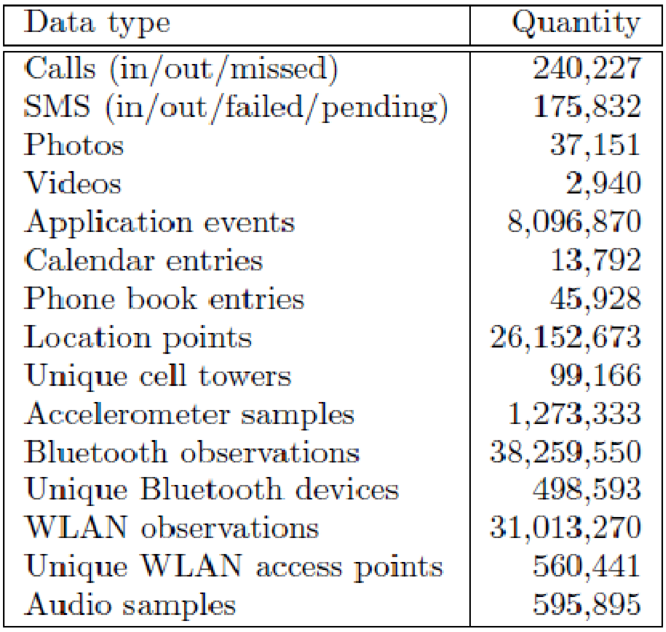
Smart phones not only help people in communication, but they are also ideal platforms for collecting mobile data. Mobile device users produce different types of data : battery charging patterns, the types of commonly used applications, where users like to go after work, which songs they listen and many more. With health applications users can track their activity levels, heart rates, their calorie intakes, etc. By using social network applications (i.e. Foursquare) companies and restaurants can learn what their costumers like or dislike. Then they can use this information to provide better service.

Using "Big Data", researchers and companies can answer many questions about users' habits (coffee consumption increase in the mornings), daily problems (traffic jam), etc. Place prediction problem is one of these questions.

One of the Mobile Data Challenges is entitled "From big smartphone data to worldwide research : the mobile data challenge" (Laurila et al., 2013). In this large scale data collection campaign (Lausanne Data Collection Campaign-LDCC) smart phone data from nearly 200 volunteers in the Lake Geneva region over 18 months is collected. The dataset consisted of precomputed attributes of participants, provided as a 100,000 - 15,000 customer-attribute matrix, and the target values for the training set. In the Mobile Data Challenge the participants were free to exploit the raw sensor data for their prediction methods.

The smart phone model used in the Lausanne Data Collection Campaign was Nokia N95 phone. Total number of participants was 185 with 62% male and 38% female distribution. The data was collected between October 2009 and March 2011.

Reality mining work "Reality mining : sensing complex social systems" (Eagle and Pentland, 2006) was focused on social systems. The data was collected for 9 months



Data type	Quantity
Calls (in/out/missed)	240,227
SMS (in/out/failed/pending)	175,832
Photos	37,151
Videos	2,940
Application events	8,096,870
Calendar entries	13,792
Phone book entries	45,928
Location points	26,152,673
Unique cell towers	99,166
Accelerometer samples	1,273,333
Bluetooth observations	38,259,550
Unique Bluetooth devices	498,593
WLAN observations	31,013,270
Unique WLAN access points	560,441
Audio samples	595,895

Figure 2.1: LDCC Main Data - Amount Data

from 100 mobile phones. Using standard Bluetooth enabled mobile phones, collected information used with different contexts and social patterns of a daily user obtained. These patterns were activity, infer relationships, identify socially significant locations, and model organizational rhythms. Modellings of individual users was done by Bluetooth and cell tower IDs. Using cell tower IDs to determine location is more visited idea then using Bluetooth device information. BlueAware application was designed to log BTIDs. It is a passive application that runs in background.

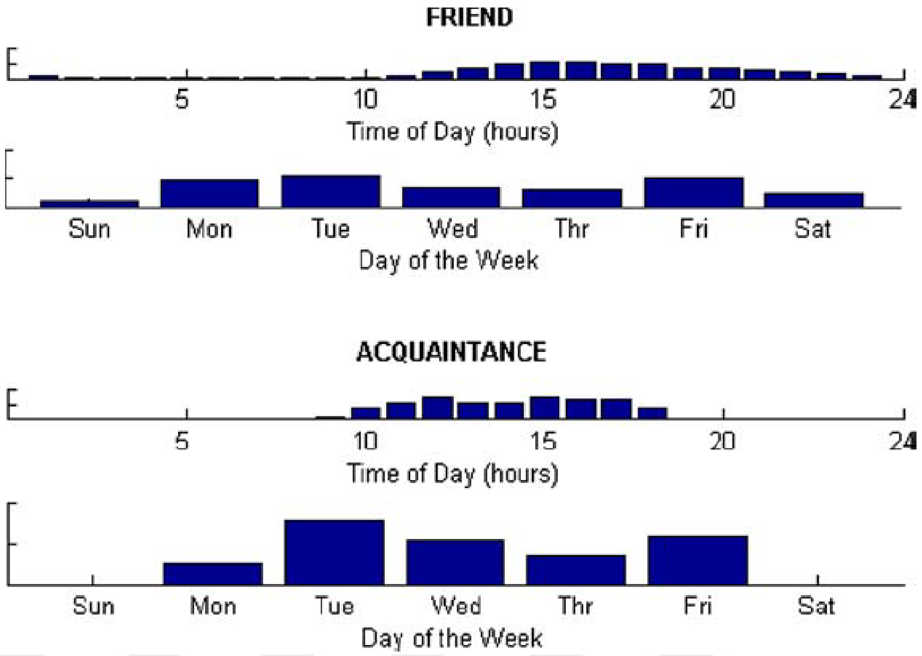


Figure 2.2: Reality Mining on Complex Social Systems

This study consisted of 100 Nokia 6600 smart phones pre-installed with several pieces of software. These are some developed by study and a context application from the University of Helsinki. The data collected are call logs, Bluetooth devices in proximity, cell tower IDs, application usage, and phone status. Data was collected during an academic year (nearly 450,000 hours). User location, communication and device behaviour.(Public, anonymous version of the dataset is available on the website http ://-reality.media.mit.edu)

This work is similar to ours it uses Time Of Day feature as a classification class. As shown in 2.2 a participants location prediction depends on the time spend between

a friend and a work place acquaintance (co-worker). A professor' s daily work routine may be similar to an other faculty member but for weekly routine a friend's behaviour is a better choice to study.

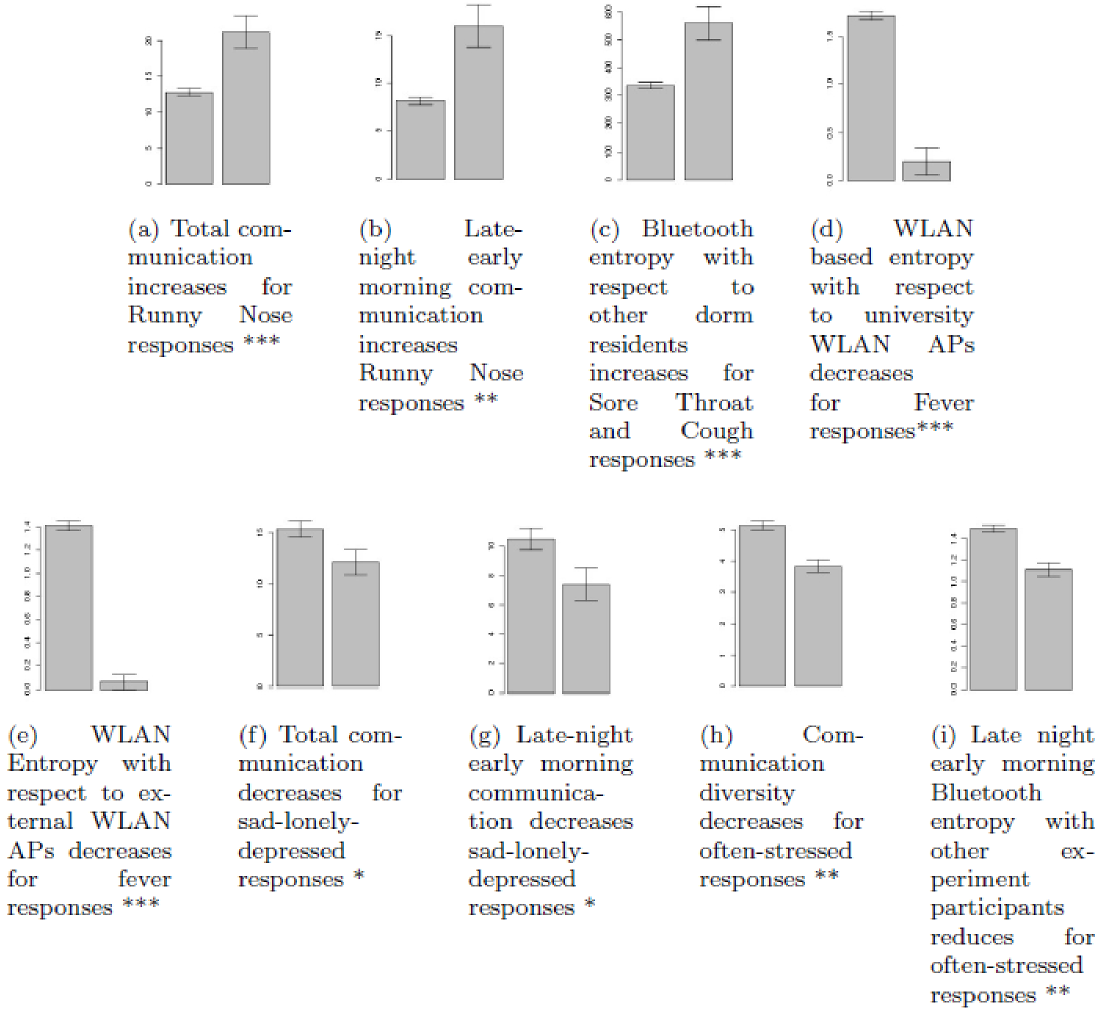
"Sensing the" health state" of a community" (Madan et al., 2012) is a study bent to the topic of health state. Individuals phone usage behaviour may change according to their mental and physical health. Madan's study predicts changes in health using phone features. Changes like cold, influenza and stress are common changes in the urban areas. Sick individuals behave differently than their healthier days. Also obesity plays a important role in the activity change. This research found that the participants with weight gain and the participants that did not gain weight have correlation with each other.

This data collection from a smart phone application was performed by 70 participants (students) at an undergraduate residence hall during an academic year (September 2008 and June 2010). Dataset consists of 3.15 million scans of Bluetooth devices, 3.63 million scans of WLAN access-points, 61,100 call data records, and 47,700 logged SMS event.

Participants took monthly self-report surveys related to their health habits, diet and exercise, weight changes, and political opinions during the presidential election campaign dated between September 2008 and June 2010, an entire academic year.

Especially during the influenza peek period (February to April 2009), participants also provided 2,994 daily symptom reports related to common colds, fever, influenza, and mental health. When the fewer and CDC defined influenza infected the participants, both WLAN entropy (university access points) and external access points showed a dramatic decrease.

This research also tracked mental health changes. If a participant was often-stressed and sad-lonely-depressed, they had shown more isolated behaviours. Total communication, late-night communication, communication diversity and late-night Bluetooth entropy decreased as well. This study also illustrates the potential of mobile phones to monitor the health status of an individual in almost real-time.



(a) Total communication increases for Runny Nose responses ***

(b) Late-night early morning communication increases Runny Nose responses **

(c) Bluetooth entropy with respect to other dorm residents increases for Sore Throat and Cough responses ***

(d) WLAN based entropy with respect to university WLAN APs decreases for Fever responses***

(e) WLAN Entropy with respect to external WLAN APs decreases for fever responses ***

(f) Total communication decreases for sad-lonely-depressed responses *

(g) Late-night early morning communication decreases sad-lonely-depressed responses *

(h) Communication diversity decreases for often-stressed responses **

(i) Late night early morning Bluetooth entropy with other experiment participants reduces for often-stressed responses **

Figure 2.3: Mandan Affect of Physical and Mental Condition to Bluetooth Entropy

2.2 Related Work on Semantic Place Prediction

2.2.1 Placer and Placer++

Placer and the continuation of this study, Placer++ algorithms are good labelling machine learning examples for our research. They have inspired us on how to name our place labels so the participants have better user experience. Placer algorithm was discussed in "Placer : semantic place labels from diary data" (Krumm and Rouhana, 2013) Their training data is collected from 87,600 place visits (from 10,372 distinct people) evaluated on 1,135,053 visits (from 124,517 distinct people). Placer algorithm identifies place labels based on the timing of visits to that place, nearby businesses, and simple demographics of the participant.

Placer algorithm uses two publicly available datasets. American Time Use Survey (ATUS) and the Puget Sound Regional Council Household Activity Survey (PSRC). ATUS dataset features are as follows :

- Age of subject in integer years
- Gender of subject
- Arrival day of week
- Arrival time of day
- Visit midpoint time of day
- Departure time of day
- Duration of visit
- Holiday (binary)
- Season of year (0,1,2,3)

PSRC participants were from the Puget Sound region in the U.S (four counties near Seattle, WA.) They have filled out a survey covering their visits over two consecutive days. The PSRC diary data includes latitude/longitude data of these visits. PSRC consists data from 86,764 trips taken by 9790 different people who gave labels to 18,888 distinct places. The arrival date and time, the duration, and the latitude/longitude of the location. PSRC is a Point of Interest dataset. It has labels like ;

- Arts & Entertainment
- Automotive & Vehicles
- Business to Business
- Computers & Technology
- Education
- Food & Dining
- Government & Community
- Health & Beauty
- Home & Family
- Legal & Finance
- Professionals & Services

- Real Estate & Construction
- Shopping
- Sports & Recreation
- Travel

Using these labels, the feature list given below is created :

- Count of each type with 50 meters
- Count of each type within 100 meters
- Count of each type within 200 meters
- Distance to nearest instance of each type

Hidden Markov Model classification algorithm is used as well as decision tree. The Placer gave an accuracy of 73.0% on this dataset.

The following Placer++ algorithm is discussed in "Placer++ : Semantic place labels beyond the visit"(Krumm et al., 2015). It was developed as a more accurate labelling method than Placer. The accuracy increased by 8.85 percentage points over the baseline of Placer. As similar to Placer, government diary data was used as dataset as well.These are same datasets, 2006 Puget Sound Regional Council (PSRC) Household Activity Survey. Also labelled data from almost 10,000 participant concentrated in a single metro area was used.

Placer++ features has two main innovations. Machine learning was used for examining sequences of of a participant's visits in order to increase labelling accuracy. They have found out that when a participant spends time at the label "college", she/he does not spend time at label "childcare". Second approach was to use other participants' label data.

Placer++ increased classification accuracy to 72.70% over the baseline accuracy of the original Placer algorithm by 8.85 %.

2.2.2 Predicting Home and Work Locations Using Public Transport Smart Card Data

The research "Predicting Home and Work Locations Using Public Transport Smart Card Data by Spectral Analysis" (Li et al., 2015) uses public transport smart card data to predict home or work place.

Singapore is used as a case study in this research. The island-wide automated fare collection system for public transportation based on payment EZ-Link card Transportation charges customers according to their travelling distance. Time and location information was collected when the card was used.

This research's data was provided by the Singapore Land Transport Authority (LTA). The dataset comprises Singapore' s three-month smart card transactions between 1 November 2011 to 31 January 2012. Time span between last alighting and consecutive boarding is less than a threshold of 45 minutes.

From the preprocessed travel records they estimate the likelihood of a bus stop or train station to be the home/work place of a commuter by considering a number of factors.

According to this research one of the most important indicators for home and work activities performed between two successive public transport journeys is its duration. However, they did not calculate the duration of the stays in the given location. They just used the time of day feature. They have assumed that a commuter will stay at home/work place alternately in weekdays and their stay duration should match his/her working hours. This method assumed that people go to work in public workdays. This result is similar to our weekday/weekend analysis explained in Chapter 3.

2.2.3 Feature engineering for semantic place prediction

"Feature engineering for semantic place prediction"(Zhu et al., 2013) study is also from Nokia Mobile Data Challenge. As classification algorithms, four different algorithms were used : Logistic Regression (LogReg), Support Vector Machines (SVM), Random Forests (RF), and Gradient Boosted Trees (GBT).

For feature selection, this study used the following features :

- Time
- Application
- Bluetooth and WLAN
- Accelerometer
- Call log
- System

— Media

2.2.4 Analyzing Location Predictability on Location-Based Social Networks

By using users check-in data, this project predicts future check-in place. This paper (Lian et al., 2014) estimates the entropy of an individual check-in trace and then leverage Fano' s inequality to transform it to predictability. They used decision trees and cross validation.

2.2.5 Mobile Persuasion : 20 Perspectives on the Future of Behavior Change

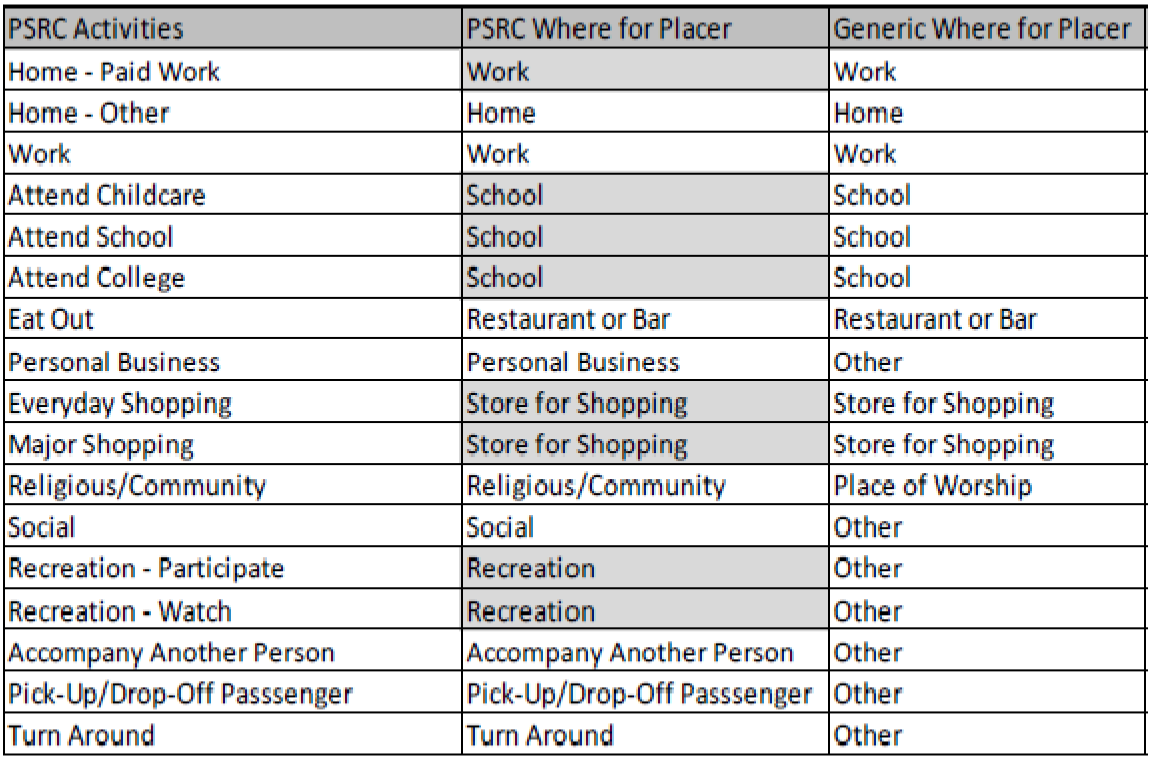
This study is another research using Mobile Computing and Location-based social networks (LBSNs). Nokia Mobile Data Challenge (Nokia MDC) provides the largest and the richest public dataset for researchers to study this problem.

Their dataset contains 112 users' mobile sensing data and each type of data may contribute discriminative power to Semantic Place Prediction. With time context alone, they were able to distinguish working places and homes accurately. After midnight people usually stay at home. In addition to time context, call log and other sensor data recorded in the user's smartphone during the user's stay at a place may also indicate the type of it. (from reference (Berchtold et al., 2010; Zhu et al., 2013))

2.2.6 Places of Our Lives

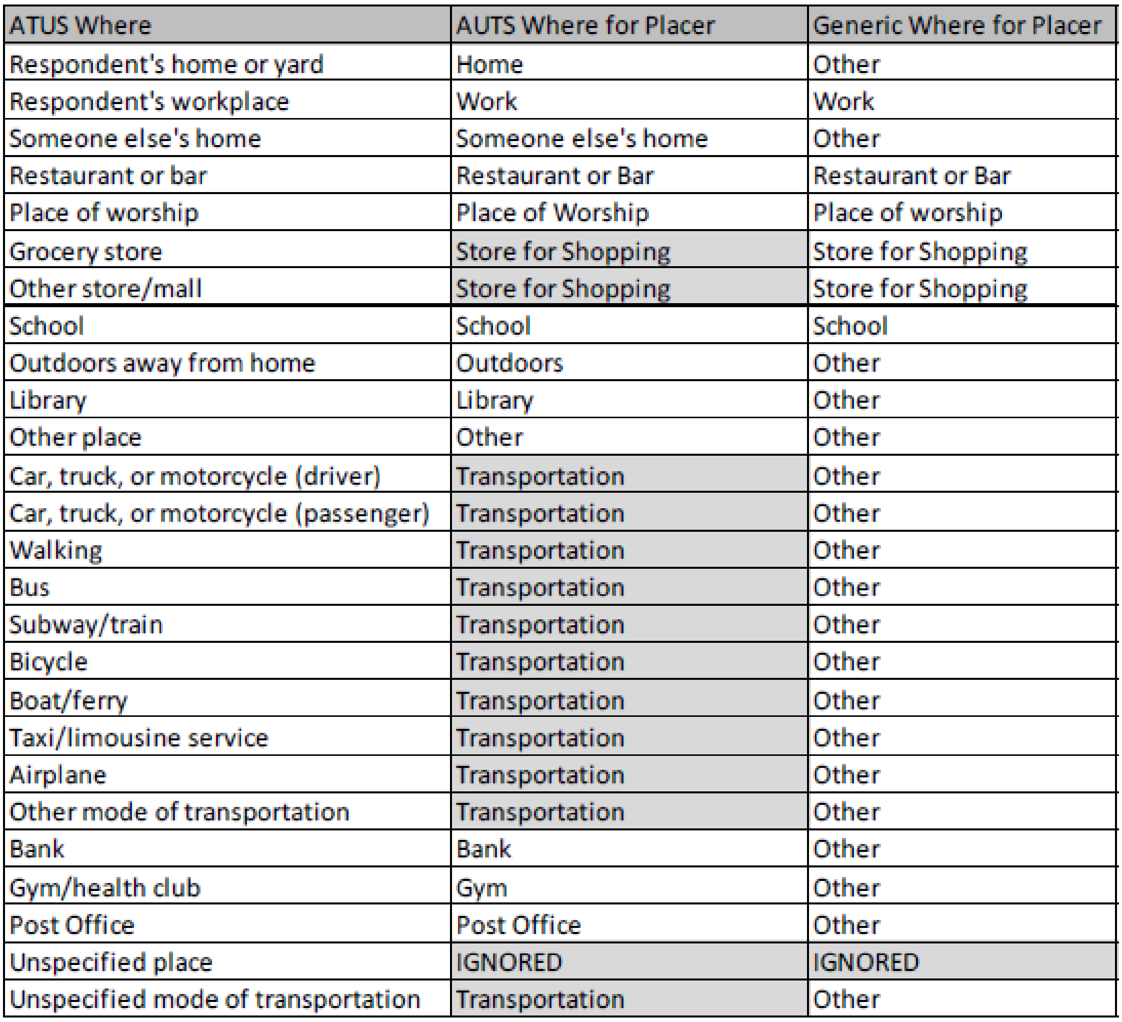
This workčitedo2014places also utilizes Nokia challenge data. However , what makes it important for our research is how they labelled the data. For our project we choose labelling data as Home, Work/School,etc as generic as possible. As mentioned in this paper, people tend to visit many undefined/non-frequent places during their daily lives. As a solution, we choose to name "Friend's/parent's Home" as a general label for all family and friends places, and use "Outdoor(Park,etc)" as for all outdoor activities like bus stops, bazaar, etc. According to the differences of activities and time we can differentiate them.

This study focuses on characterization of real-life place visiting patterns from smart phone data and automatic place labelling in a location privacy sensitive setting. They reveal findings regarding both regularly and novelty trends in the visiting patterns. Considering the problem of place labelling with 10 place categories, they have shown that frequently visited places could be recognized reliably (over 80%).



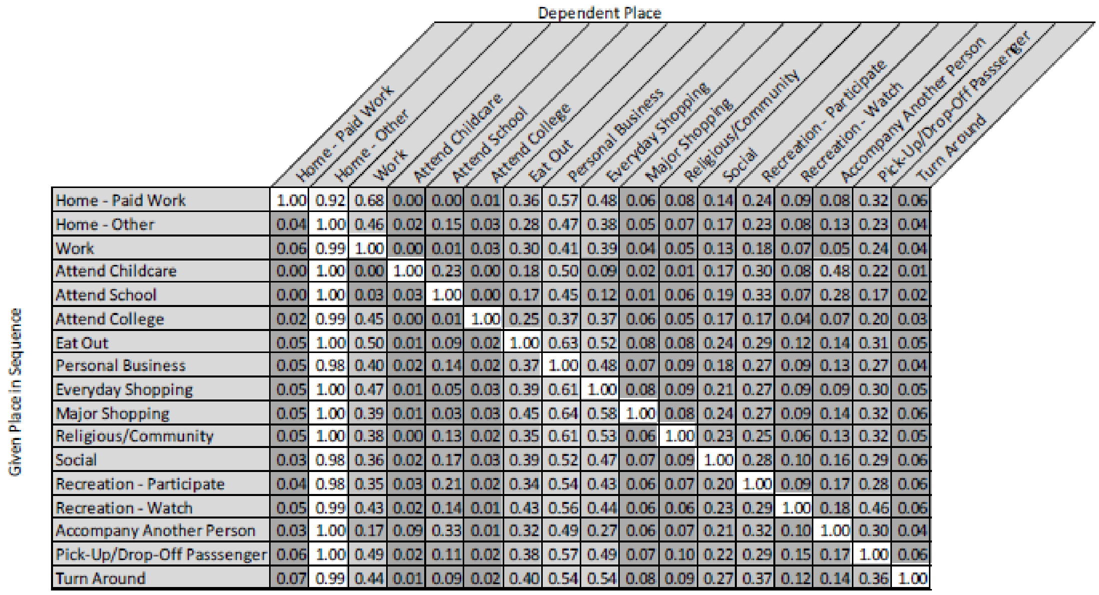
PSRC Activities	PSRC Where for Placer	Generic Where for Placer
Home - Paid Work	Work	Work
Home - Other	Home	Home
Work	Work	Work
Attend Childcare	School	School
Attend School	School	School
Attend College	School	School
Eat Out	Restaurant or Bar	Restaurant or Bar
Personal Business	Personal Business	Other
Everyday Shopping	Store for Shopping	Store for Shopping
Major Shopping	Store for Shopping	Store for Shopping
Religious/Community	Religious/Community	Place of Worship
Social	Social	Other
Recreation - Participate	Recreation	Other
Recreation - Watch	Recreation	Other
Accompany Another Person	Accompany Another Person	Other
Pick-Up/Drop-Off Passsenger	Pick-Up/Drop-Off Passsenger	Other
Turn Around	Turn Around	Other

Figure 2.4: Placer PSRC Table



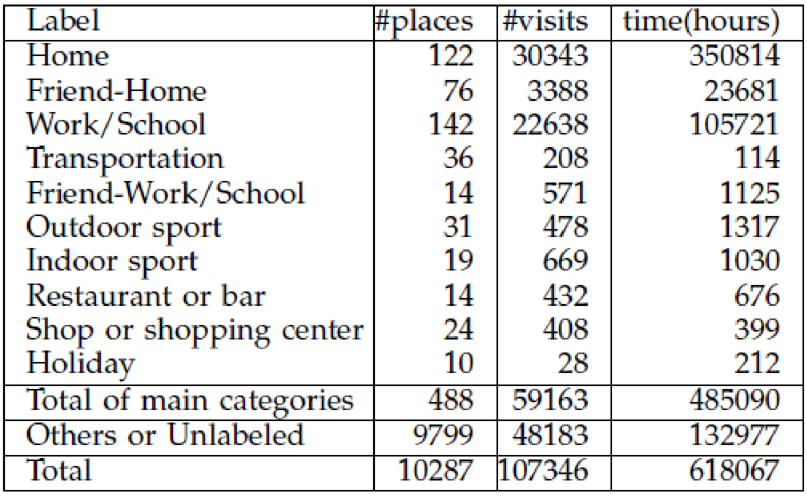
ATUS Where	AUTS Where for Placer	Generic Where for Placer
Respondent's home or yard	Home	Other
Respondent's workplace	Work	Work
Someone else's home	Someone else's home	Other
Restaurant or bar	Restaurant or Bar	Restaurant or Bar
Place of worship	Place of Worship	Place of worship
Grocery store	Store for Shopping	Store for Shopping
Other store/mall	Store for Shopping	Store for Shopping
School	School	School
Outdoors away from home	Outdoors	Other
Library	Library	Other
Other place	Other	Other
Car, truck, or motorcycle (driver)	Transportation	Other
Car, truck, or motorcycle (passenger)	Transportation	Other
Walking	Transportation	Other
Bus	Transportation	Other
Subway/train	Transportation	Other
Bicycle	Transportation	Other
Boat/ferry	Transportation	Other
Taxi/limousine service	Transportation	Other
Airplane	Transportation	Other
Other mode of transportation	Transportation	Other
Bank	Bank	Other
Gym/health club	Gym	Other
Post Office	Post Office	Other
Unspecified place	IGNORED	IGNORED
Unspecified mode of transportation	Transportation	Other

Figure 2.5: Placer Atus Label Table



	Dependent Place																
Given Place in Sequence	Home - Paid Work	Home - Other	Work	Attend Childcare	Attend School	Attend College	Eat Out	Personal Business	Everyday Shopping	Major Shopping	Religious/Community	Social	Recreation - Participate	Recreation - Watch	Accompany Another Person	Pick-Up/Drop-Off Passsenger	Turn Around
Home - Paid Work	1.00	0.92	0.68	0.00	0.00	0.01	0.36	0.57	0.48	0.06	0.08	0.14	0.24	0.09	0.08	0.32	0.06
Home - Other	0.04	1.00	0.46	0.02	0.15	0.03	0.28	0.47	0.38	0.05	0.07	0.17	0.23	0.08	0.13	0.23	0.04
Work	0.06	0.99	1.00	0.00	0.01	0.03	0.30	0.41	0.39	0.04	0.05	0.13	0.18	0.07	0.05	0.24	0.04
Attend Childcare	0.00	1.00	0.00	1.00	0.23	0.00	0.18	0.50	0.09	0.02	0.01	0.17	0.30	0.08	0.48	0.22	0.01
Attend School	0.00	1.00	0.03	0.03	1.00	0.00	0.17	0.45	0.12	0.01	0.06	0.19	0.33	0.07	0.28	0.17	0.02
Attend College	0.02	0.99	0.45	0.00	0.01	1.00	0.25	0.37	0.37	0.06	0.05	0.17	0.17	0.04	0.07	0.20	0.03
Eat Out	0.05	1.00	0.50	0.01	0.09	0.02	1.00	0.63	0.52	0.08	0.08	0.24	0.29	0.12	0.14	0.31	0.05
Personal Business	0.05	0.98	0.40	0.02	0.14	0.02	0.37	1.00	0.48	0.07	0.09	0.18	0.27	0.09	0.13	0.27	0.04
Everyday Shopping	0.05	1.00	0.47	0.01	0.05	0.03	0.39	0.61	1.00	0.08	0.09	0.21	0.27	0.09	0.09	0.30	0.05
Major Shopping	0.05	1.00	0.39	0.01	0.03	0.03	0.45	0.64	0.58	1.00	0.08	0.24	0.27	0.09	0.14	0.32	0.06
Religious/Community	0.05	1.00	0.38	0.00	0.13	0.02	0.35	0.61	0.53	0.06	1.00	0.23	0.25	0.06	0.13	0.32	0.05
Social	0.03	0.98	0.36	0.02	0.17	0.03	0.39	0.52	0.47	0.07	0.09	1.00	0.28	0.10	0.16	0.29	0.06
Recreation - Participate	0.04	0.98	0.35	0.03	0.21	0.02	0.34	0.54	0.43	0.06	0.07	0.20	1.00	0.09	0.17	0.28	0.06
Recreation - Watch	0.05	0.99	0.43	0.02	0.14	0.01	0.43	0.56	0.44	0.06	0.06	0.23	0.29	1.00	0.18	0.46	0.06
Accompany Another Person	0.03	1.00	0.17	0.09	0.33	0.01	0.32	0.49	0.27	0.06	0.07	0.21	0.32	0.10	1.00	0.30	0.04
Pick-Up/Drop-Off Passsenger	0.06	1.00	0.49	0.02	0.11	0.02	0.38	0.57	0.49	0.07	0.10	0.22	0.29	0.15	0.17	1.00	0.06
Turn Around	0.07	0.99	0.44	0.01	0.09	0.02	0.40	0.54	0.54	0.08	0.09	0.27	0.37	0.12	0.14	0.36	1.00

Figure 2.6: Placer++ Probability to Attend Other Places



Label	#places	#visits	time(hours)
Home	122	30343	350814
Friend-Home	76	3388	23681
Work/School	142	22638	105721
Transportation	36	208	114
Friend-Work/School	14	571	1125
Outdoor sport	31	478	1317
Indoor sport	19	669	1030
Restaurant or bar	14	432	676
Shop or shopping center	24	408	399
Holiday	10	28	212
Total of main categories	488	59163	485090
Others or Unlabeled	9799	48183	132977
Total	10287	107346	618067

Figure 2.7: 10 Main Category of The Places of Our Lives

Unlike our data verifying process they gave a survey at the end of their process. Together with annotation, they have also used feedback on the meaning and the quality of some of the discovered places (5 top frequent places and 3 infrequent places). Highly positive feedback was obtained for the set of frequent places, with 95% of them were confirmed by the users. One fifth of the set of infrequent places were not remembered by people.

This paper shows that it is not easy to recognize the semantic meaning for a large number of places in our lives if the actual physical location is not known. Frequently visited places such as home, work or the home of a friend can be reliably recognized using only location-sensitive smart phone data. For instance, we combined Family and friend' s home label together.

Since we collected our own data, we start with an early observation stage, then applied these necessary changes into small group of newly collected data as a second stage. Last data collection stage came with small modifications on our code and more participants for collection. In Tables 2.1 and 2.2, we provide a comparison with related studies.

In our study, we explored a different methodology than these related studies. Unlike our study, the other studies were not focused on the relation with activity and location. We try to establish a connection between the places visited during re participants daily routine and try to guess these places with participants' performed activities besides other features. Do participants sit during transportation ? Can we identify the places with activities ? This way we study the effect of the Activity Features on predicting places. Time features and communication features were also utilized in similar studies, however, among the phone related features, while application and media usage were utilized, battery level, headset usage were not explored.

An other difference was the collection of the data set. The preliminary and month-long data collection let us to study the different features. During preliminary data collection, we could adapt our methodology to return better prediction rates. In our main data collection, we included different sets of features such as phone features, as well. As mentioned, previous studies mainly utilized the available datasets, such as Nokia dataset (Laurila et al., 2012). Since we utilized the ARService application with real-time activity classification capability with high accuracy, we could gather information about the participants' activities which was not available in previous datasets.

Paper Refe-rence	Device	Data Type	Number of Participants	Data Collection Duration	Features Used	Classification Algorithms	Dataset	MDC Participant	Result
(Laurila et al., 2013)	Nokia N95	GPS, cell tower IDs, call records, WiFi, geographic coordinates, app log events, Bluetooth records	200 people	18 months	Time of Day, Mobility, WiFi, Application Usage, Bluetooth, etc.	KNN and SVM	LDCC	Yes	Predicts from old data
(Eagle and Pent-land, 2006)	Phone	Bluetooth, cell tower IDs	100 phones	9 months	Time of Day	(BlueAware, Bluedar)	MIT self made	No	Predicts social systems
(Madan et al., 2012)	Smart Phone	Bluetooth devices, WLAN access points, SMS events, call records	70 people	1 academic year	Alters, Events and Times, Time of Day		Self collected	No	Phone usage behaviour changes according to their mental and physical health.
(Krumm and Rou-hana, 2013)	Smart Phone	GPS	1,135,053 visits from 124,517 people. 1,135,053 ATUS visits and 87,600 PSRC visits.	PSRC : two continuous day.	Age of subject in integer years Gender of subject Arrival day of week Arrival time of day Visit midpoint time of day Departure time of day Duration of visit Holiday (binary)Season of year (0,1,2,3) Count of each type with 50 meters Count of each type within 100 meters Count of each type within 200 meters Distance to nearest instance of each type	Placer, Decision Tree, Hidden Markov Model	American Time Use Survey (ATUS) and the Puget Sound Regional Council Household Activity Survey (PSRC)	No	Both diary studies gave classification accuracies of around 0.73. Sampled GPS at a 10-second interval.
(Krumm et al., 2015)	Smart Phone	GPS	PSRC 47,060 unique person/place pairs. Labelled data from almost 10,000 people concentrated in a single metro area.	PSRC : two continuous day.	Age, gender, surrounding business and timing features.	Placer++	American Time Use Survey (ATUS) and the Puget Sound Regional Council Household Activity Survey (PSRC)	No	7.39% of the time the same participant gave a different label to the same place

Table 2.1: Comparison with Related Work

Paper Reference	Device	Data Type	Number of Participants	Data Collection Duration	Features Used	Classification Algorithms	Dataset	MDC Participant	Result
(Li et al., 2015)	Smart Card	Public Transport Fare Card	Singapore Public Transport	1 November 2011 to 31 January 2012	Time and location	-	Singapore Land Transport Authority (LTA)	No	This method assumed that people go to work in public workdays.
(Do and Gatica-Perez, 2014)	N95	GPS,WiFi, geographic coordinates, app log events, and Bluetooth records, among several other sensor types	114 participants	18 month	Mobility, WiFi, Applcation Usage, Bluetooth, Time Features	Tukey-Kramer method together with ANOVA	LDCC	Yes	The study suggests that personalized recommendation could be a plausible approach.
(Berchtold et al., 2010)	N95	GPS,WiFi, geographic coordinates, app log events, and Bluetooth records, among several other sensor types	114 participants	18 month	Mobility, WiFi, Application Usage, Bluetooth, Time Features		LDCC	Yes	Calculates the duration of transportation and stay between Home- Work labels.
(Zhu et al., 2013)	N95	GPS, WiFi, geographic coordinates, app log events, and Bluetooth records, among several other sensor types	114 participants	18 month	Time, Application, Bluetooth WLan, Accelerometer, Call log, System, Media and Bag-of-words (BoW)	Logistic Regression (LogReg), Support Vector Machines (SVM), Random Forests (RF), and Gradient Boosted Trees (GBT)	LDCC	Yes	They have find that the features conditioned on fine-granularity time intervals are very helpful for place category classification and our conditional feature construction effectively constructs these features in a principled way.

Table 2.2: Related Paper Comparison

2.3 Tools

2.3.1 WEKA

WEKA (Waikato Environment for Knowledge Analysis) (Hall et al., 2009) is a Java based machine learning software developed by Waikato University (New Zealand) and it is used for applying classification and feature selection algorithms on our data. Weka can be used not only with desktop computers but also with smart phones. Users can choose training and test data easily and modify many variables through options. User may use .arff, .csv, .data, etc. file types with Weka. We mostly use .csv files as our file type.

Weka Software is easy to use, it gives users classification and testing options. As seen in the figures 2.8, 2.9 we can choose a filter and apply selected data attributes. As

an example we use "Normalize" filter with whole attributes set. For each attributes (Time, Mean, sdX, sdY, sdZ, Var, FFT1,FFT2, FFT3, FFT4, PITCH, ROLL, Ori, Act) a graph has been drawn. 2.10

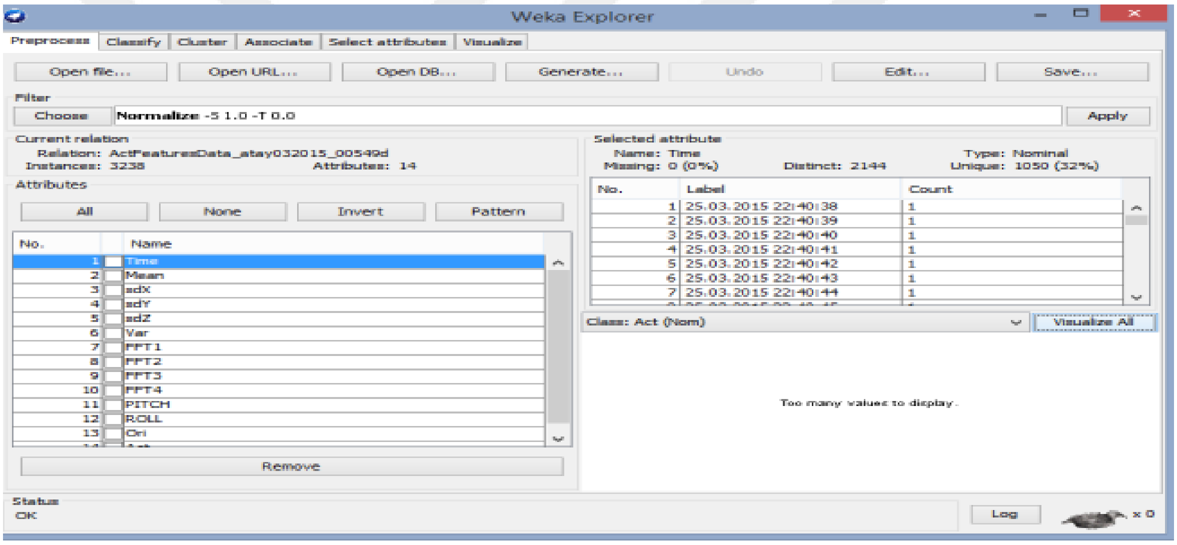
2.3.2 ARService

ARService is a mobile data collection application developed in (Coskun, 2014) at Galatasaray University. This application has been used for collecting the location-activity dataset used in this thesis. In the data collection phase, volunteers are students and faculty members of Galatasaray University. This application collects various data such as GPS, WiFi, accelerometer, running applications on the smart phone, battery level. In this project we use machine learning algorithms to solve place prediction problem with using these pre-collected data.

ARService has been developed for Android platform (Coskun, 2014). The system is composed of client-side mobile application and the server side where data collected from users are uploaded. In order to save battery, in ARService attention is paid for data uploading and sleep-wake time sharing. ARService automatically sends collected data packages to the server, only when the WiFi connection is established and when sensors are not sampled.

WEKA
The University of Waikato
Weka logo, featuring weka, a bird endemic to New Zealand

Figure 2.8: WEKA Software



Filter
Normalize -S 1.0 -T 0.0

Current relation
Relation: ActFeaturesData_atay032015_00549d
Instances: 3238
Attributes: 14

Attributes

No.	Name
1	Time
2	Mean
3	sdX
4	sdY
5	sdZ
6	Var
7	FFT1
8	FFT2
9	FFT3
10	FFT4
11	PITCH
12	ROLL
13	Ori

Selected attribute
Name: Time
Missing: 0 (0%)
Distinct: 2144
Type: Nominal
Unique: 1050 (32%)

No.	Label	Count
1	25.03.2015 22:40:38	1
2	25.03.2015 22:40:39	1
3	25.03.2015 22:40:40	1
4	25.03.2015 22:40:41	1
5	25.03.2015 22:40:42	1
6	25.03.2015 22:40:43	1
7	25.03.2015 22:40:44	1

Class: Act (Nom)

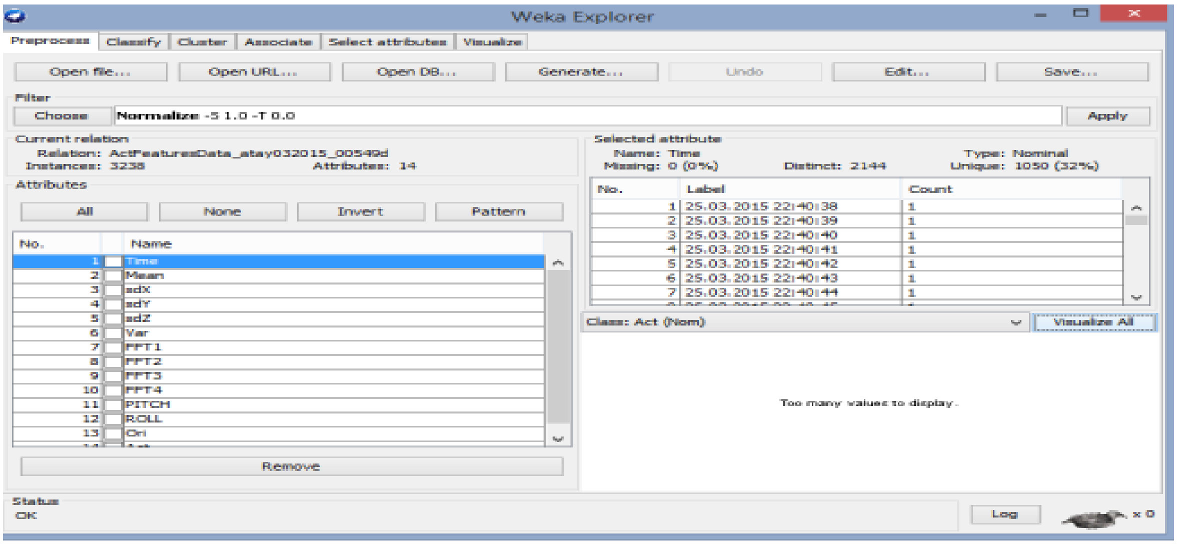
Too many values to display.

Status
OK

Figure 2.9: Example for ActFeatureData using Normalize filter with all whole attributes selected.

Login Screen, shown in Table 2.3, lets the users to enter their name, age, height, and weight with password they choose. After creating their account, users can see their user page and track their process with the given graphs on the server side.

In every 15 minutes, an automatic pop-up window is shown to the user. In the window, the possible predicted activity of the last 2 minutes is presented. The screen shows the most probable two activities recognized by the online classifier. If the predicted activities are not correct, the user can input the correct activity. If the user does not interact with this window, it closes in the next 45 seconds. For this thesis, ARService was modified to ask for user's location as well. After the activity recognition popup window, another popup window appears and asks the user about the place she/he



Filter: Normalize -S 1.0 -T 0.0

Current relation
Relation: ActFeaturesData_atay032015_00549d
Instances: 3238 Attributes: 14

Attributes

No.		Name
1	□	Time
2	□	Mean
3	□	sdX
4	□	sdY
5	□	sdZ
6	□	Var
7	□	FFT1
8	□	FFT2
9	□	FFT3
10	□	FFT4
11	□	PITCH
12	□	ROLL
13	□	Ori

Selected attribute
Name: Time Type: Nominal
Missing: 0 (0%) Distinct: 2144 Unique: 1050 (32%)

No.	Label	Count
1	25.03.2015 22:40:38	1
2	25.03.2015 22:40:39	1
3	25.03.2015 22:40:40	1
4	25.03.2015 22:40:41	1
5	25.03.2015 22:40:42	1
6	25.03.2015 22:40:43	1
7	25.03.2015 22:40:44	1

Class: Act (Nom)

Too many values to display.

Status
OK

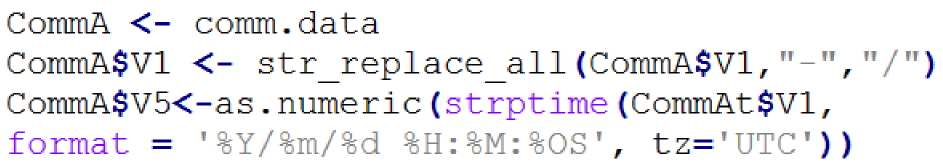
Figure 2.10: Our Results.

has been. The user can select a place from the presented list. This data is used as ground truth information for place recognition. This popup window is also presented in Table 3.1. The list of place labels are : Home, Friend's/parent's home, Work/School,

On the Road (Transportation), Outdoor (Park, etc.), Canteen/Restaurant/Cafe/Bar, Mall/Shop, Other, [GSÜ] Classroom/Lab, [GSÜ] Canteen/garden, [GSÜ] Library. The last three places are specific to our university campus whereas the others are selected according to the most popular places visited during daily life. Similar list of places were adopted in related studies as well (Do and Gatica-Perez, 2014).

Using “Manual Mark up” page, a user can also enter the place label manually. The same list of places is presented to the user as in the place popup window, but the user can also enter new place names that are not on the list.

Mission option lets the user collect new annotation data to increase the precision of activity prediction rate. For the 2 minutes, the smart phone collects data and matches the activity tags with the given tags. User can choose between "Walking", "Running", "Transportation", "Stairs", "StandingStill",and "Sitting" activities. At the second option, user can choose a phone location as "Pants Pocket", "Jacket Pocket", "Chest Pocket", "Backpack", "Messenger-Shoulder Bag", "Handbag", "Belt", or "Hand" places. When the mission duration ends, ARService ask user "if it is a valid mission?". This way the correction of the matching done by user at first.



```
CommA <- comm.data
CommA$V1 <- str_replace_all(CommA$V1,"-","/")
CommA$V5<-as.numeric(strptime(CommAt$V1,
format = '%Y/%m/%d %H:%M:%OS', tz='UTC'))
```

Figure 2.11: R Code Example

2.3.3 R Programming Language

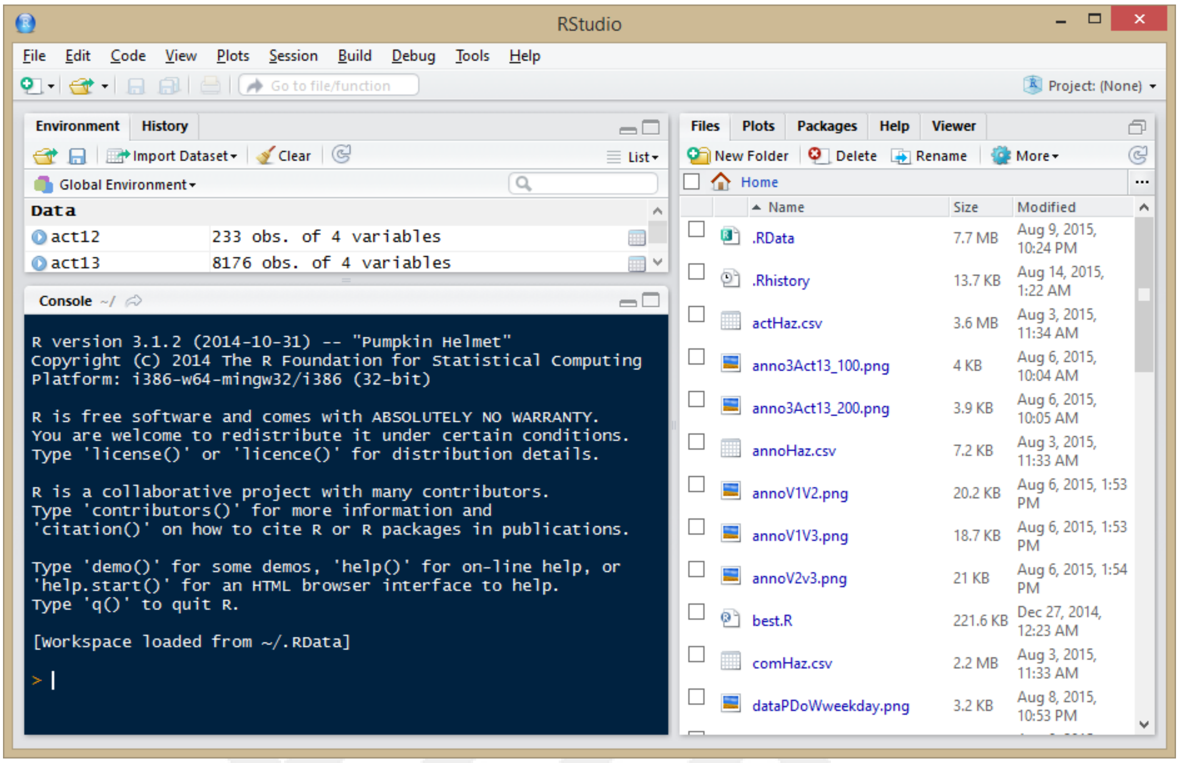
R is an implementation of the S programming language combined with lexical scoping semantics inspired by Scheme. S was created by John Chambers while at Bell Labs. There are some important differences, but much of the code written for S runs unaltered. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team, of which Chambers is a member. R is named partly after the first names of the first two R authors and partly as a play on the name of S. (RStudio Team, 2015).

R language is used in data processing stages through this thesis. R can be used in not only mobile devices but also with web programs. It is light. And R languages updates frequently by the people use it. Some features of the R language :

- Big Data Analytics
- Machine Learning Analytics
- Graphics and Visualization
- Data Mining and Machine Learning
- Statistical Methodology
- and many more.

R language does not need compile code before using it. Since it is an open source code, users can find specific package for their specific needs. From analysing big data to creating a web page you can find different options. In this thesis we have used diplyr, tidyr and stringr String manipulation libraries, as well as lubridate date- time manipulation library. In our first and second stages of data processing we face with date- time conversion problems. the example code below is a code snippet of how we solved it.

We used R and R Studio for preprocessing the data and in combining different types of data, namely communication, phone, activity data (explained in Section 3).


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Project: (None)
Environment History
Import Dataset
Clear
List
Global Environment
Console ~/
Files Plots Packages Help Viewer
New Folder Delete Rename More
Home

Data

act12	233 obs. of 4 variables
act13	8176 obs. of 4 variables

```
R version 3.1.2 (2014-10-31) -- "Pumpkin Helmet"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

> |
```

Name	Size	Modified
.RData	7.7 MB	Aug 9, 2015, 10:24 PM
.Rhistory	13.7 KB	Aug 14, 2015, 1:22 AM
actHaz.csv	3.6 MB	Aug 3, 2015, 11:34 AM
anno3Act13_100.png	4 KB	Aug 6, 2015, 10:04 AM
anno3Act13_200.png	3.9 KB	Aug 6, 2015, 10:05 AM
annoHaz.csv	7.2 KB	Aug 3, 2015, 11:33 AM
annoV1V2.png	20.2 KB	Aug 6, 2015, 1:53 PM
annoV1V3.png	18.7 KB	Aug 6, 2015, 1:53 PM
annoV2v3.png	21 KB	Aug 6, 2015, 1:54 PM
best.R	221.6 KB	Dec 27, 2014, 12:23 AM
comHaz.csv	2.2 MB	Aug 3, 2015, 11:33 AM
dataPDoWweekday.png	3.2 KB	Aug 8, 2015, 10:53 PM

Figure 2.12: R Studio User Interface

2.3.4 R Studio

R Studio is a free software with main aim of ease of use. Many of R libraries are deve by students, thus they may not be dependable. By using R Studio users can keep track of updates and changes. Also one can be found many library packages for some specific subject, i.e. graphics. However between these packages contains different attributes for same things.

In figure 2.12, we can see R Studio user interface. On the left, Environment section

shows the data and value and this can be used without redefinition. Console section is the place we code. On the right bottom corner users can see plots, help information, etc. ActFeature data of a user collected during March 2015 is presented in Figure 2.13.

In the plotted graphic we draw sdX, sdY, sdZ features.

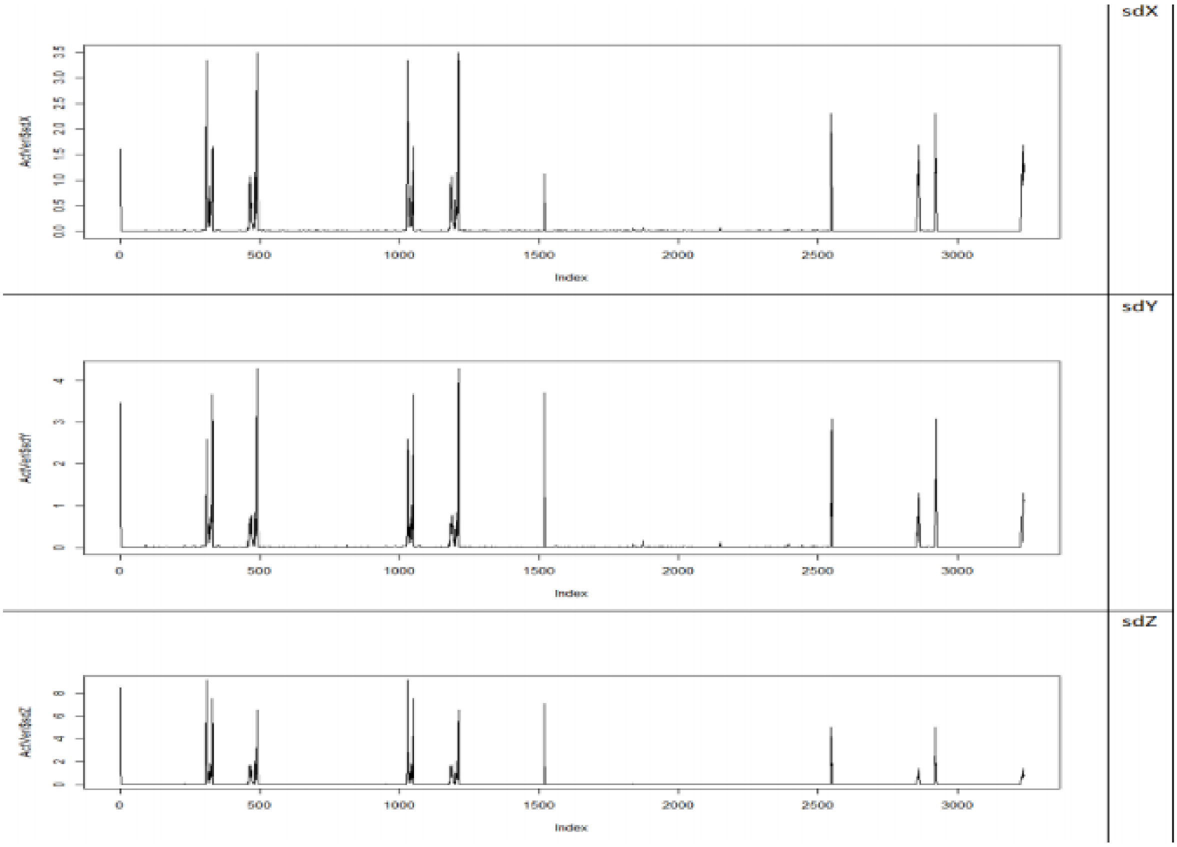
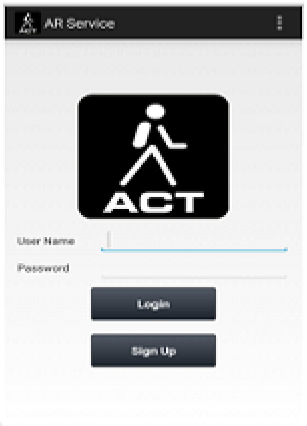
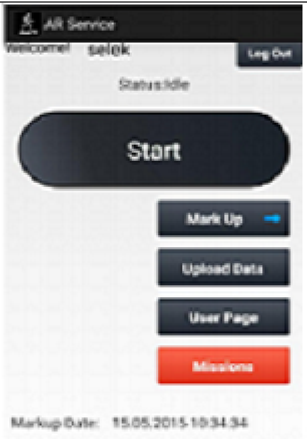
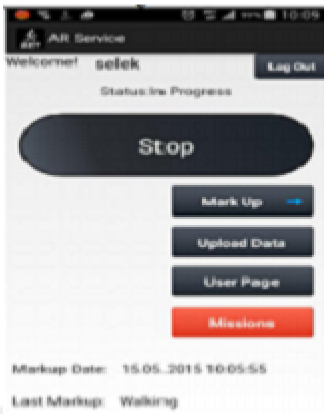
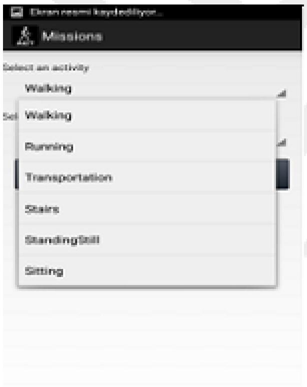
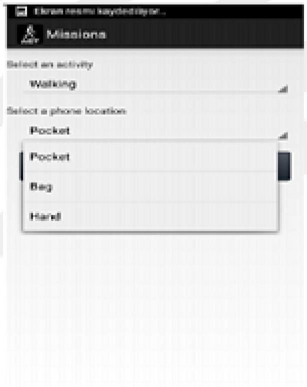
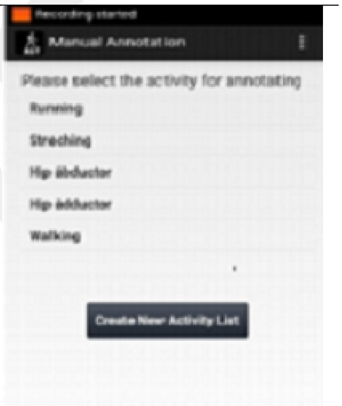


Figure 2.13: Tables for ActFeatureData created with RStudio.

Login Screen	User Activity	Manual Mark up
 AR Service ACT User Name Password Login Sign Up	 AR Service Welcome! selek Log Out Status:Idle Start Mark Up Upload Data User Page Missions Markup Date: 15.05.2015 10:34:34	 AR Service Welcome! selek Log Out Status:In Progress Stop Mark Up Upload Data User Page Missions Markup Date: 15.05.2015 10:05:55 Last Markup: Walking

Selection Type	Selection Device	Manual Annotation
 Ekran resmi kaydediliyor... Missions Select an activity Walking Walking Running Transportation Stairs StandingStill Sitting	 Ekran resmi kaydediliyor... Missions Select an activity Walking Select a phone location Pocket Pocket Bag Hand	 Recording started Manual Annotation Please select the activity for annotating Running Streching Hip abductor Hip adductor Walking Create New Activity List

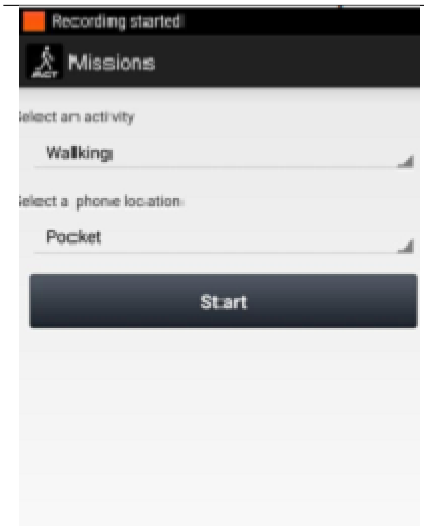
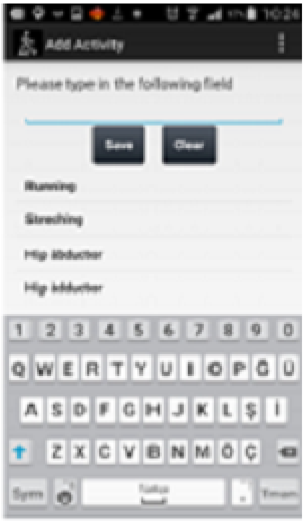
Mission Option	Create Mark up
 Recording started Missions Select an activity Walking Select a phone location Pocket Start	 Add Activity Please type in the following field Save Clear Running Streching Hip abductor Hip adductor

Table 2.3: ARService Smart Phone Application

3 SEMANTIC PLACE PREDICTION METHODOLOGY

In this section, we explain the methodology for place prediction. We elaborate on the details of each step, from data collection to classification.

3.1 Place Prediction Framework

In figure 3.1, we present the steps taken for place prediction. Initially, raw data is collected from mobile phone. Next, different tables from raw data are merged together and features are extracted. Finally, the training and testing steps start. In the following sections, we explain the details of each step.

3.1.1 Data Collection and Preprocessing

Collecting the data and preprocessing it were the most important steps of our study. There were three stages of data collection. First stage consists of November 2015 to January 2016. Four people and six smart phones contribute to the collection. Smart phone models were Samsung Galaxy S3 Mini and Samsung Galaxy S5. Two of the participants used two phones of different models at the same time period. Participants were one male faculty member and one female faculty member, one female undergraduate student and a female master student. In this initial step, we tested our methodology with limited number of users and improved it for a bigger dataset.

Second stage consists more than 20 users. Same four users still contributed at this stage. But this time students of Galatasaray University (17 participants) joined the data collection. The data collection took place from April 1 to May 2, 2016, for about 31 days. We think that the main difference in places between an undergraduate student and a master student is their frequency of attending to school, while the difference between an undergraduate student and a faculty member is mostly about their out of working/school hours.

Our raw data contains Activity Information, Annotation Data, Communication Infor-

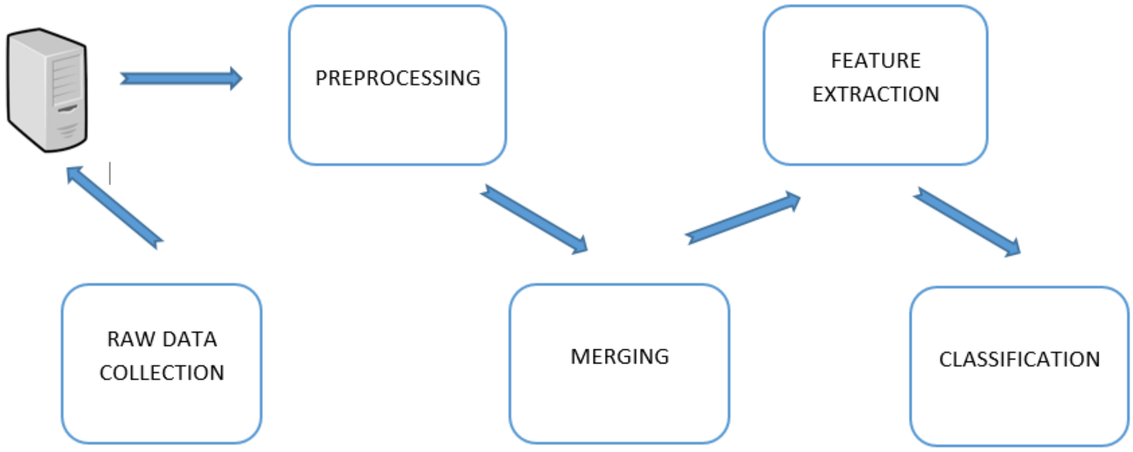


Figure 3.1: Steps of Place Prediction

mation, Phone Data and Location Information. However, we choose to not use Location Information for classification or feature selection due to privacy reasons. Our aim is predicting location without any location information, such as GPS data.

ARService creates four main data files. These four main are files called CommInfo, PhoneAct, Annotation, and ActFeatures files. These four data tables were the source of the our raw data. In the CommInfo table, data related to the wireless network usage is preserved. The columns are named as : "Time","Feature","Field",and "Extra". Detailed description of this data is given below.

— Time :This column collects time data in the format of Year(four digits) - month(two digits) - day(two digits) hour : minutes : seconds. three digits of milliseconds (YYYY-MM-DD HH :MM :SS.mmm)

— Feature : Column collects data about what kind of action, the user's smart phone performs. Feature consists of "WiFi", "location","data" and "calls" subgroups.

— Field : Depending on which Feature stored, this column can get different values. Since this column directly effects "Extra" column values as well, the possible results of "Extra" column it was discussed here, too.

— Extra : Column values correlate with Field and Feature columns.

The detailed description of these groups given below.

Feature as WiFi : The WiFi option shown when wireless connection was establish.

— "network state","CONNECTED" : The network connection was established.

— "connected to"," @ xx :xx :xx :xx :xx :xx" gives the id of the connected device.

— "state","enabled" : Displays that the connection is enabled.

Feature as calls : Extra column will display the state of call. For example ("service state","normal"), ("idle","no activity").

Feature as location :

- "provider","cell tower ids" what kind of provider the smart phone scanned.
- "cell tower id","xxxxxxx" the seven digit id of the connected cell tower.
- "location area code","xxxxx" the location are code for the connected cell tower.

Feature as data :

- "connection state","disconnected" or "connected".
- "activity" field value we can get "None", "sending data" and "sending & receiving data" activity types of our smart phone into the Extra column.

CommInfo table is our base table for merging. All the merging was done by sorting CommInfo time frame. During our observation state and second stage we classified all the repeating data as ell. This repeating caused by measurement and collection time interval differences. Same data tables stored different data by milliseconds, while others collected by seconds. To a better and faster classification we removed this duplicated data by hanging our processing script.

The Annotation table includes the activity and place labels for the data collected. The column values are listed as "Time", "PossibleAnnotations", and "Annotation".

- Time : Collects time data in the format of four digits year - month - day hour : minutes : seconds. three digits of milliseconds (YYYY-MM-DD HH :MM :SS.mmm) It is same with CommInfo data table format.
- "PossibleAnnotations" : It is calculated by the ARService activity recognition algorithm depending on the most possible Annotation type the participant might be done at the two minutes time frame. If the participant sits for a while then walks around, then the ARService provides a popup selection menu for the user. "StandingStill(0,956)/Sitting(0,044)" is an example for the stored "PossibleAnnotations" data. Here with a 0.956 to 0.044 possibility that the participant was standing still instead of sitting during recording. In the modified version of ARService where we added location popups, this column is used to identify place labels from the activity labels. If this column has "-" then it means user is manually entering labels. If this column has "Location" as the input then this is the answer a user has given to the location popup window.

- "Annotation" : The real annotation. It is either recorded answer of the participant to popup selection menu, or the manual mark up from the participant.

The ActFeatures data table consists of "Time", "Mean", "sdX", "sdY", "sdZ", "Var", "FFT1", "FFT2", "FFT3", "FFT4", "PITCH", "ROLL", "ORIEN", "Act" and "IsSent" column set :

- "Time" column stores the current date in year - month - day hour : minutes : seconds. three digits of milliseconds format (YYYY-MM-DD HH :MM :SS.mmm).
- "Mean" stores the mean of sdX, sdY and sdZ.
- "sdX" Android smart phones can collect standard deviation of the activity in the x axis.
- "sdY" Android smart phones can collect standard deviation of the activity in the y axis.
- "sdZ" Android smart phones can collect standard deviation of the activity in the z axis.
- "Var" Variable column.
- "FFT1" is Fast Fourier Transforms of data.
- "FFT2" is Fast Fourier Transforms of data.
- "FFT3" is Fast Fourier Transforms of data.
- "FFT4" is Fast Fourier Transforms of data.
- "PITCH" ? smart phones can detect pitch, roll and azumute of its location. Pitch collects the x axis of movement. Pitch sensor gives positive values when the z-axis (azule) moves toward the y-axis (roll).
- "ROLL" Roll data collects the y axis data(-90 to 90) increasing as the device moves clockwise.
- "ORIEN" is the orientation data of smart phone. It is defined as a combination of three angular quantities azimuth, pitch, and roll data.
- "Act" ARService classification algorithm generates this column. "Trans", "Walking", "StandingStill", "Sitting", "Running" and "Stairs" labels were our results.

The last data table we used for classification is "PhoneAct" table which includes phone usage details. This file has columns called "Time", "Feature", "Field", and "Extra". As you can see the column names are same with CommInfo data table. And their connection with each other is quite similar. Extra column variables depends on Field column variables, while Field column variables changes according to Feature variables. Again Field and Extra column has variable groups responds to same Feature variable. These groups and how each column is sorted can be described as ;

- "Time" Collects time data in the format of four digits year - month - day hour : minutes : seconds. three digits of milliseconds (YYYY-MM-DD HH :MM :SS.mmm).
- "Feature" This column can get "Application", "Screen" and "Battery" values.
- "Field" When "Application" data collected in "Feature" column the responding value group is the name of the last active application. For "Application" our Field column can have the following options.

  "Feature" column as a "Screen", then Field and Extra columns can get the :

  - "turned on" formed when the device detects the screen is turned on.
  - "turned off" formed when the device detects the screen is turned off.

  options.

  Feature as "Battery Level"would get The "Battery" as Field value. The remaining battery level would be stored in Extra column.
- "Extra" column depends on "Feature" and "Field" columns. For "Application" value the application name (ie. "AR Service") would be followed by "x". However if the next value is "Running Applications", then the "Extra" column will shows the number of the running applications at that time. When the "Screen" value recorded, regardless of "Field" value, the "Extra" would be a "x". Again, for the "Battery Level", the remaining battery level by 100 percentage would be given.
- "IsSent" depending the connection between smart phone and server this column can get 0 for failure and 1 for successful connection.

ARService also provides LocationInfo file, however, our main goal is predicting the next place of a user without any GPS information. Thus LocationInfo table was not used during any classification steps.

As our data set, we can group it into three different stages. First four participants collected during all these three stages. We will name them as our base group. Base group consists of 2 faculty members, and 2 students (1 was a undergraduate student, while the other was a master student). While one of the faculty members is male, rest of the base group are female.

Age difference can cause daily routine differences as well. While a young (for this study below 25 years/ undergraduate student) might prefer hanging out at late night hours, while older participants prefer to spend this time at home due to their responsibilities (kids at home, early working hours, etc.) Another difference of the age groups can be shown as their activity level. "WalkingCountR" and "RunningCountR" values might be higher for young individuals. However, activity levels also depends the participants life style.

Battery level consumption is a major problem in mobile data collection applications. One of the set backs of smart phone usage is battery consumption. The battery cells only store certain amount of energy and if sensors are continuously sampled and data is continously sent to the server on ARService application, battery may deplete within a few hours. In order to minimize battery consumption, we used a duty-cycling method. 5 minute cycles were used and sensors were sampled for 2 minutes and in the remaining 3 minutes, sampling was stopped. This duty cycling method lets us to use the phone at least from morning to evening for an average user.

ARService also provides LocationInfo file, however, our main goal is predicting the next place of the user without any GPS information. Thus LocationInfo table was not used during any classification steps.

3.1.2 Merging Files and Feature Extraction

As mentioned in the previous section, we have four files extracted from ARService : Phone data, activity data, communication data and the annotation data. Annotation data includes the tags that were input by the users. These tags include information both about the performed activities and the places visited. At the end of processing them, we have Time, Communication, Activity and Phone Features. Our data mining process continues with grouping these features during classification to reach maximum efficiency.

3.1.2.1 Time Features

Each of our raw data tables has a date column. But each of them has different format. In order to be able merge these tables we first created an Epoch Feature. Epoch Feature is the numeric representation of current date. It is also called Unix time as well. Epoch time the total seconds spend from Thursday, 1 January 1970 or 00 :00 :00. Since milliseconds important for ordering and sorting our raw data epoch time used. At the first stage with our minimal (4 participant data) this feature used in the classification. However at the second and third stage of our study we only use epoch time feature as processing/merging our raw data, not in the classification phase.

Day Feature lets us be able to see what are the daily routines of participants. Do our behaviours change according to days of the week ? Almost no participant has 9 to 17 daily works so, these changes can be seen. We have hour based routines Hour Feature added. At 8 pm people tend to be on the road to work/school while at the 20 o'clock they tend to be at home. Hour Feature helps us to track daily routine of participants. Working class and students tend to attend the morning communities, so they were more active during these hours.

ToD or Time of Day Feature lets us segment a day into four parts. Mornings (between 6 pm to 11 pm), Afternoon (between 12 to 17), Evening (between 18 to 23), and Night(between 24 to 5). At Night period there tends to be less tags than others, because between 24 to 5 people with daily jobs/lessons prefer to sleep. Any inconsistent tagging can be detected easily.

Weekday.weekend Feature is included for tracking peoples daily life changes. While people tend to go to Work/School during early hours of weekdays ; people like to take leisure time on weekends. In total we have 5 features extracted from the Time data.

3.1.2.2 Communication Features

When our smart phones connect with wireless communication devices (cell towers, WiFi hot-spots, routers, etc.) ARSevice collects the device ids and number of connections. Communication Features are created using these data.

The part of Communication Features are, "UniqueBSID" and "UniqueSSID". "UniqueBSID" stores the total number of the unique cell towers id' s. Normally a smart phone scans at periodic time intervals for cell towers. This way connection will not interrupt even if the owner of the smart phone is mobile. Android lets the developers collect these cell tower' s MAC address. However it causes privacy issues. So, while grouping each tower id' s with the location label ; we choose to calculate the total number of different cell towers that our device scanned.

For example CommInfo data file stores "2016-03-29 11 :48 :58.664","location","provider","cell tower ids" "2016-03-29 11 :48 :58.668","location","cell tower id","10553xxxx" "2016-03-29 11 :48 :58.690","location","location area code","64xxx" group at 29 March 2016 at 11 :48. As you can see in 36 milliseconds smart phone (Samsung Galaxy S5) scanned it's environment and found a cell tower, then stored it's id and local area code. (x's used as numbers to provide privacy)

"UniqueSSID" collects the wireless network devices that our smart phone scanned. Again similar to "UniqueBSID", we only calculate the number of the different devices between time intervals. As an example the scanning and connection process start at row "2016-03-29 13 :28 :54.413", "Wifi", "scan result", " @ yy :yy :yy :yy :yy :yy-GSU" and ends with "2016-03-29 13 :30 :36.465", "Wifi" ,"scan result", " @ xx :xx :xx :xx :xx :xx-GSU" "2016-03-29 13 :33 :56.136", "Wifi", "network state", "CONNECTED" "2016-03-29 13 :33 :56.140", "Wifi", "connected to", " @ xx :xx :xx :xx :xx :xx" takes less then 2 seconds. All this time the smart phone scanned 67 different wireless devices.

The reason for choosing unique BSID and SSID numbers instead of the exact IDs, is the protection of anonymity. Since SSID' s and BSID' s collect the name of the scanned devices, one can track the movement of the participant. For example, in the campus of Galatasaray University access points have ”GSU” label in their names. Additionally, some users chose to name their access points like ”home”, ”work”, etc. By using total count of these devices ; we can observe the change of the number of the scanned devices through participant's daily routes. Some of the semantic place prediction works collect the GPS data, social network markings, etc. to predict a place. This may cause privacy violation. In our research we try to prove, that one can make a high rate prediction without disclosing location information.

3.1.2.3 Activity Features

For activity tracking, ARService provides an automatic recognition and questioning. Every two minutes a pop-up lets the participants choose from prediction list. And if the participant has another action, they can chose "Neither" and add manually. Walking, Sitting, StandingStill, Running, Stairs, Transportation are our main/ automatic tags.

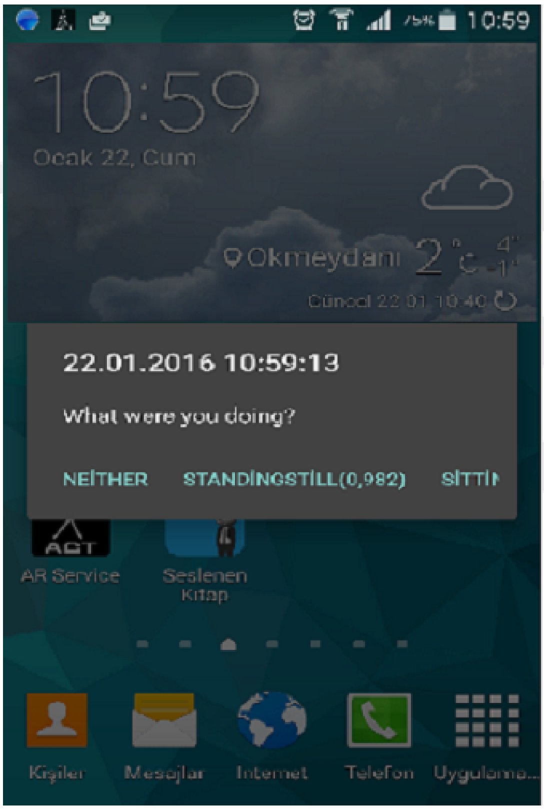
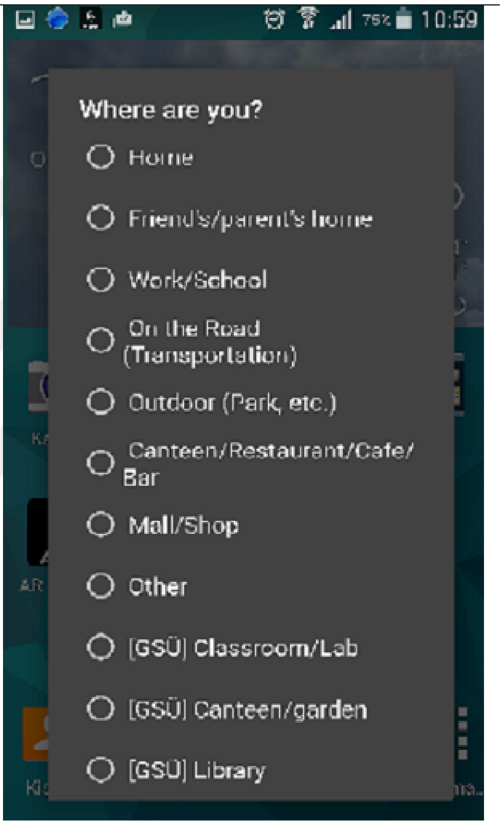
Prediction Percentage Window	Location Tag Window
 22.01.2016 10:59:13 What were you doing? 10:59 Ocak 22, Cum Okmeydanı 2°C NEİTHER STANDİNGSTİLL(0,982) SİTTİ AR Service Seslenen Kitap Kişiler Mesajlar Internet Telefon Uygulama...	 Where are you? ○ Home ○ Friend's/parent's home ○ Work/School ○ On the Road (Transportation) ○ Outdoor (Park, etc.) ○ Canteen/Restaurant/Cafe/ Bar ○ Mall/Shop ○ Other ○ [GSÜ] Classroom/Lab ○ [GSÜ] Canteen/garden ○ [GSÜ] Library 10:59

Table 3.1: ARService smart phone label options

Activity grouping of ARService gives us 8 different types of activities. These actions are, "Sitting", "Walking", "Transportation", "StandingStill", "Mobile", "Running", "Stairs" and "Standing". Activity Feature has the corresponding "WalkingCount", "TransportationCount", "StandingStillCount", "MobileCount", "RunningCount", "StairsCount" and "StandingCount" For the duration of the data collected from participant we counted the number of times these 8 activities occurred. Let us say for 15 minutes data collection, our participant may have walked for 10 minutes, then decided to sit on a bank for the remaining 5 minutes. In the Activity Feature all the occurrence are calculated. Then this result would be normalized with the total occurence of all the activities. If the participant had 200 walking instances, 150 sitting instances and 150 transportation instances ; the results would be 0.40000 for WalkingCount, 0.30000 for SittingCount and TransCount.

Unfortunately "Stairs" activity has a classification problem in ARService, in the data we collected "Stairs" returned zero value for every participant.

3.1.2.4 Phone Features

Phone data table contains the smart phone device usage information. Detailed definition of these data are given below.

1. Number of total running application gives us both background and foreground running applications. ARService is always accounted during our measurement.
2. Screen turned on or off. When do the participants lean to turn on screen of their smart phone ? Is it during morning commuting ? Or the participants prefer to open it during working hours ? Reason to turning on or off the screen may differ. Participant may wish to check the time, read their incoming message, or tried to skip the song they were listening. As smart phone's collected features we have choose to use "Head set", "Battery Level", "Running Application Numbers". Head set feature, the device controls if the user is using any head set gear. If the user is using it(Bluetooth sets,head phones for listening music, etc.) then our "HeadSetCount" feature will get 1, else it will be a zero. Battery level is one of the most changing feature of a smart phone. The more active the user gets, the faster the battery levels drop. Smart phone users tend to carry their recharge devices with them. Some using cables and use landlines to charge. So this group must be in a building, rather it is "Home", "Work/School" or any other location. However if the user has a portable charging device, then user can charge on the go. We wondered if the charging habit help us to determine the prediction of the location. "BatteryLevel" Feature collects the current battery level of the phone as percentage. As we have round the Activity Features "StllCount" to "StillRCount", we have decided to round the battery level and create a threshold for it. "BatteryThreshold" values calculated as :
   - 1 : for 0-20 percentage of remaining battery.
   - 2 : for 21-40 percentage of remaining battery.
   - 3 : for 41-60 percentage of remaining battery.
   - 4 : for 61-80 percentage of remaining battery.

— 5 : for over 80 percentage of remaining battery.

"RunningAppCount" feature is created with running applications feature of smart phone. When the "Feature" column gets "Running Application Number" we can get the extract number from "Extra" column. The number of application used in a smart phone can change depending on how active is the user. Some applications like touch screen, wireless scanning, etc. are default application of the device. While texting, photo sharing, social media account application usage can increase with the user's activity. We wondered if the age plays any factor on the smart phone usage, or just the participant preferred different applications to use.

3.1.3 Feature Selection

Feature selection is one of the key points of machine learning. Our feature set contains four main feature groups (Time, Communication, Phone and Activity) with various subgroups.

The most common problem we have faced was matching the manual labels. Some participants choose to use Turkish characters for example [GSÜ]Classroom/Lab instead of [GSU]Classroom/Lab writing "yürüyüş" instead of "Walking". Each of these mismatches had to be changed on preprocessing the raw data. One of the common errors was using white space characters. Especially for the labels with more than one word (Outdoor (Park, etc.)) white space character was entered during manual mark up.

The necessary steps we need to take in order to have good results can be listed as ;

1. Over fitting : It is a regression model occurs when you attempt to estimate too many parameters from a sample that is too small. In our first stage of data collection we faced overfitting. In order to overcome this problem we tried removing double variables, rounding our results to two digit, etc.
2. Feature selection : To achieve high results in relevant results, we need to choose between the features of our data. Is the Day Feature more relevant then Battery-Levels Feature ? Or is the Hour Feature returns better results then Time of Day Feature ?
3. Attribute Feature selection : Selecting the attribute with the highest information gain.

Attribute selection with Weka allows the users to choose between different search algorithms. Correlation based Feature Selection (CFS) algorithm is one of the most commonly used attribute selection algorithms.

1. Best First : According to official Weka tutorial, this method searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility. Setting the number of consecutive non-improving nodes allowed controls the level of backtracking done. Best first may start with the empty set of attributes and search forward, or start with the full set of attributes and search backward, or start at any point and search in both directions.

2. Linear Forward Selection : is an extension of Best First. It takes a restricted number of k attributes into account. Fixed-set selects a fixed number k of attributes, whereas k is increased in each step when fixed-width is selected. The search uses either the initial ordering to select the top k attributes, or performs a ranking (with the same evalutator the search uses later on). The search direction can be forward, or floating forward selection (with opitional backward search steps). (Gütlein et al., 2009)

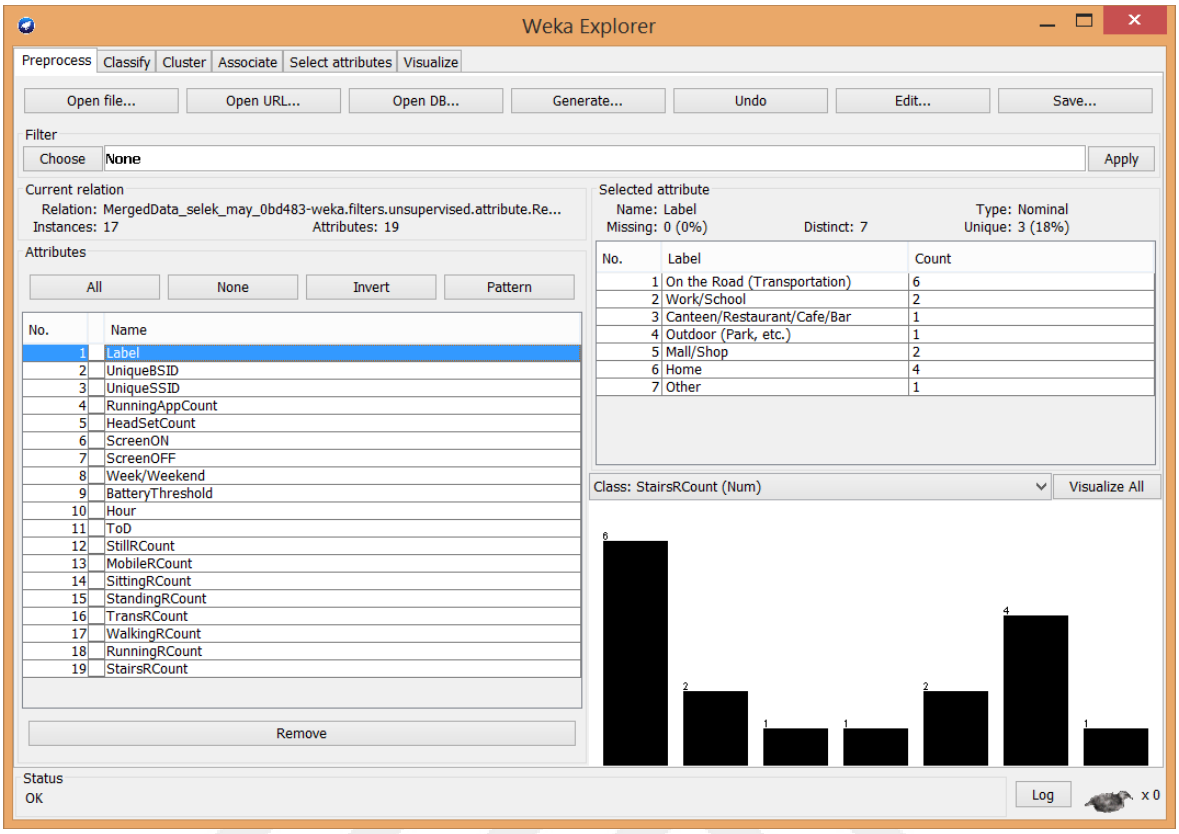
3. RankSearch : Ranks attributes by their individual evaluations. Use in conjunction with attribute evaluators (ReliefF, GainRatio, Entropy etc).

All the features used in our classification looks like Figure 3.2. As you can see in the right side of the window is our main class "Label" has 7 different variables in the given data table (April data of one participant).

According to Mark Hill's thesis (Hall, 1999) CFS was evaluated by experiments on artificial and natural datasets. Three machine learning algorithms were used : C4.5 (a decision tree learner), IB1 (an instance based learner), and Naive Bayes.

1. CfssubsetEval : Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.Subsets of features that are highly correlated with the class while having low intercorrelation are preferred. (Hall, 1999)

2. Relief Attribute Eval : Evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class. Can operate on both discrete and continuous class data. (Kira and Rendell, 1992)



Filter: None

Current relation
Relation: MergedData_selek_may_0bd483-weka.filters.unsupervised.attribute.Re...
Instances: 17 Attributes: 19

Selected attribute
Name: Label Type: Nominal
Missing: 0 (0%) Distinct: 7 Unique: 3 (18%)

Attributes

No.	Name
1	Label
2	UniqueBSID
3	UniqueSSID
4	RunningAppCount
5	HeadSetCount
6	ScreenON
7	ScreenOFF
8	Week/Weekend
9	BatteryThreshold
10	Hour
11	ToD
12	StillRCount
13	MobileRCount
14	SittingRCount
15	StandingRCount
16	TransRCount
17	WalkingRCount
18	RunningRCount
19	StairsRCount

No.	Label	Count
1	On the Road (Transportation)	6
2	Work/School	2
3	Canteen/Restaurant/Cafe/Bar	1
4	Outdoor (Park, etc.)	1
5	Mall/Shop	2
6	Home	4
7	Other	1

Class: StairsRCount (Num)

Status
OK

Figure 3.2: Processing Our Merged Data with Weka

3. Classifier Subset Eval : Evaluates attribute subsets on training data or a separate hold out testing set. It uses a classifier to estimate the 'merit' of a set of attributes.

3.1.4 Training and Classification

After preparing collected data into the homogeneous data structure, we could start to train the dataset. Testing and training data set partition needs to be done carefully. If the training is done in a haphazard way, unpredictable results may be received. Especially data over fitting is a common problem.

1. Cross Validation is one of the most popular model estimation methods of machine learning. Cross validation holds back part of the data set as a test model and use the rest of the data as training data set.

   K-fold cross validation divides data set into equal sized k parts. Within these subsets only one subset used to validate testing model. Rest of the (k-1) subsets used as training data set. K-fold cross validation helps user to avoid over fitting. The models used during data processing are given below.

   (a) 10 Fold Cross Validation : It is the most preferred k-fold cross validation

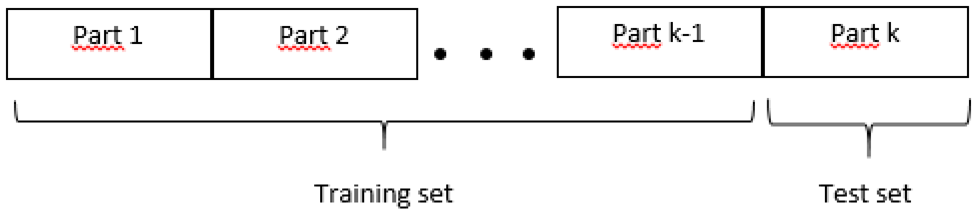


Figure 3.3: K-Fold Cross Validation

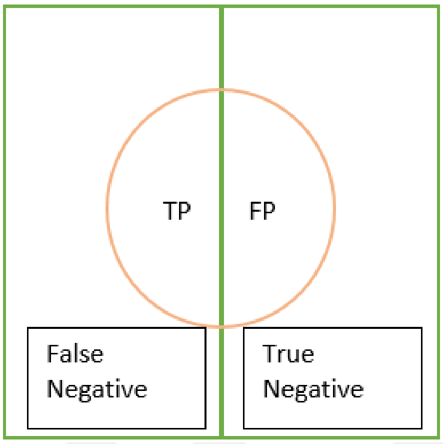


Figure 3.4: True False Positive Venn Diagram

method. The dataset will be divided to 10 equal parts.

(b) 7 Fold Cross Validation : The dataset will be divided to 7 equal parts. Following results obtained with the 7-fold .

(c) 3 Fold Cross Validation : The dataset will be divided to 3 equal parts. Following results obtained with the 3-fold.

2. Leave-one-subject out

As the performance metrics, we used the following metrics :

1. True Positive : These values are the rightly guessed ones as can be seen in the figure 3.4 Relevant documents/data and retrieved documents/data correlation.

2. False Positive : These values guessed as positive even they are actually false results.

3. Recall : True Positive result numbers' ratio to all relevant documents/data. (True Positive / True Positive + False Positive)

4. Precision : True Positive result numbers' ratio to all retrieved documents/data.

$$F - Measure = \frac{2 * precision * recall}{precision + recall} \quad Accuracy = \frac{\sum_{i=1}^{N} TP_i}{Total}$$

$$Precision = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TI_i} \quad Recall = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TT_i}$$

Figure 3.5: Performance Evaluation Metrics Equations

(True Positive / True Positive + False Negative)

5. Accuracy : All True result numbers' ratio to total population.

6. F-measure : It combines precision and recall is the harmonic mean of precision and recall, the traditional F-measure. We can show it as; F-measure = 2 * ((Precision.Recall) / Precision + Recall)) or = 2*TP / (2*TP) + FP + FN

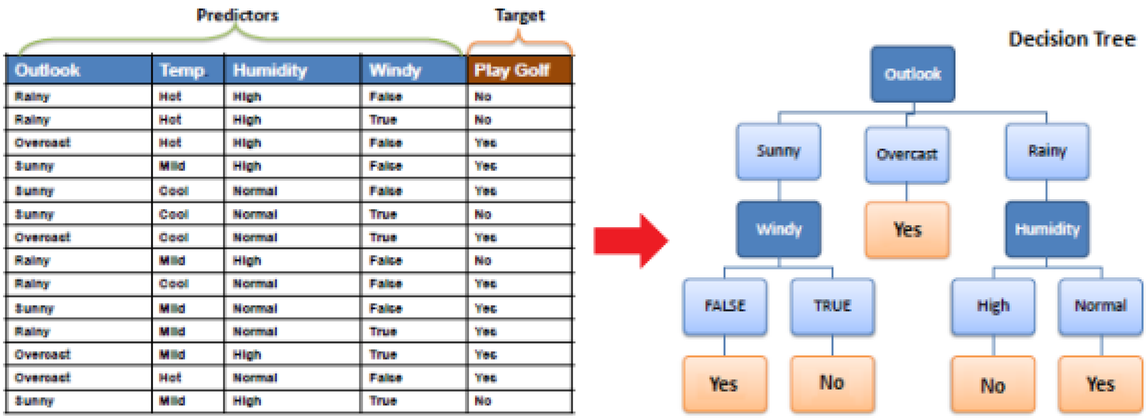
For classification the performance evaluations of our models are made by using precision, recall, F-measure, or accuracy. We also use confusion matrices. In the confusion matrix, the rows show the ground truth labels as provided by a human annotator, while the columns show the labels inferred by the model. The diagonal of the matrix shows true positives (TP). The sum of each row provides us the total ground truth of each label (TT). Lastly, the sum of each column gives us total of inferred labels (TI). The precision and the recall are separately calculated for each class and then the average is taken over all existing classes.

3.1.5 Classification Algorithms

In this section we explain the classification algorithms that are used in the thesis.

3.1.5.1 Naive Bayes

This classifier is based on Bayes' theorem with independence assumptions between predictors. This model is easy to build, with no complicated iterative parameter estimation which makes it particularly useful for very large datasets. It is quicker than discriminative models. Which means user needs fewer training data. Since Naive Bayes is simple yet efficient, this model is highly in demand.



Predictors				Target
Outlook	Temp.	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Figure 3.6: Decision Tree Example

3.1.5.2 Decision Trees

A decision tree builds classification and regression models in the shape of tree. It contains decision nodes and leaf nodes. 3.6 Decision node has least two branches. The

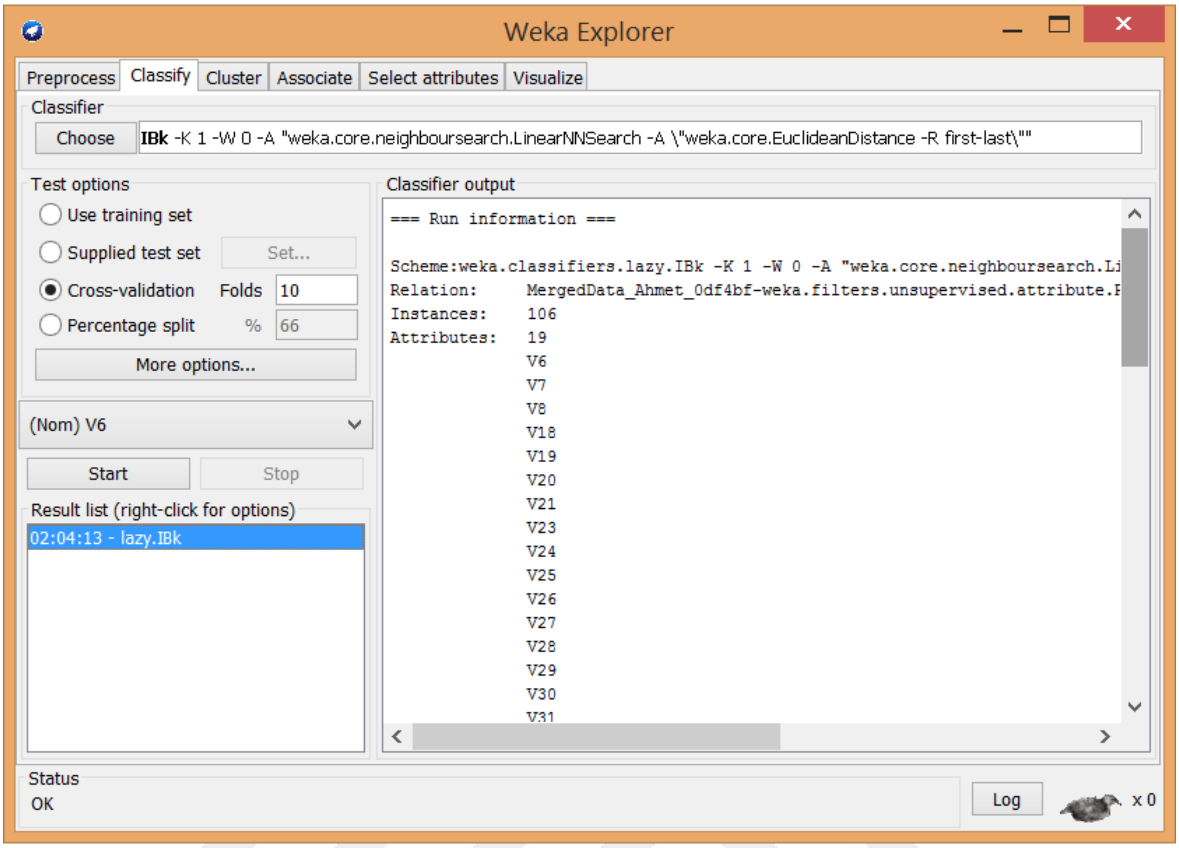
top most decision node in a tree called root node. Leaf node represents a classification or decision. We can use decision trees for both categorical and numerical data. It is easy to interpret and explain. User needs not to worry about outliers or if the data is linearly separable.

3.1.5.3 Random Forest

A Random Forest is an ensemble of decision tree classification algorithm. In Random Forest algorithm ; each decision tree will be constructed by randomized subset training data. By using same training set ; random forest tries to reduce the variance. It uses out-of-bag error as an estimate of generalized error.

3.1.5.4 KNN

KNN or the K Nearest Neighbour classification algorithm is the one of the most popular classification algorithms. In this classification, the output is a class membership. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours. K is mostly a small integer. If k = 1, then the object is simply assigned to the class of that single nearest neighbour.



Classifier

IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) V6

Result list (right-click for options)

02:04:13 - lazy.IBk

Classifier output

```
=== Run information ===

Scheme:weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.Li
Relation:     MergedData_Ahmet_0df4bf-weka.filters.unsupervised.attribute.F
Instances:    106
Attributes:   19
              V6
              V7
              V8
              V18
              V19
              V20
              V21
              V23
              V24
              V25
              V26
              V27
              V28
              V29
              V30
              V31
```

Status

OK

Figure 3.7: KNN on Weka

KNN also called a lazy algorithm. In the Weka classification algorithms, KNN sorted under lazy section, in the name of IBk.

4 PERFORMANCE RESULTS

In this section, we first give the results of semantic place prediction on the preliminary dataset collected by 4 participants. Next, we visualize the feature type-place relationship on the full dataset collected from 17 participants for one month in Section 4.2.

Afterwards, we present the results of prediction on the full dataset using cross validation in Section 4.3. We analyze the performance with different feature sets. Next,

we focus on feature selection and analyze which features provide better results in the classification in Section 4.4. Finally, in Section 4.5.2 we explore the performance in

a person-independent manner and provide the results with the leave-one-subject-out validation technique.

4.1 Preliminary Results

The stage 1 and 2 are considered as the preliminary stage of the data collection. The preliminary stage of data collection started in December 2015 with four participants. Two faculty members, a master student and a undergraduate student of Computer Engineering at Galatasaray University. Data collection ended in March 2016. Stage 1 consists of December 2015 data ; while stage 2 consists of February-March 2016 data.

At the first stage, we have observed the data storage differences between smart phones (Samsung Galaxy S5 and Samsung Galaxy S3 Mini). The questions that we focus on in the first stage were : How the labels should be arranged, should we give the participants a pre-defined label list or let them write as they wish, how could we improve the data storage problem and how should we divide time into subgroups ? These problems and any other problems occurred during these two stages let us improve our data collection and processing phases.

The most common problem we have faced was matching the manual labels entered by the users. Some participants choose to use Turkish characters for example [GSÜ]Classroom/Lab instead of [GSU]Classroom/Lab, writing "yürüyüş" instead of

"Walking". Each of these mismatches had to be changed on preprocessing the raw data. One of the common errors was using white space characters. Especially for the labels with more then one word (Outdoor (Park, etc.)) white space character entered during manual mark up.

At the first stage of data collection ; participants may create their own tags as well. This created a grouping and classification problem. Some participants created specialized tags like "taking a walk", "boat", "bicycle", etc. on the other hand, other participants just used the "On the Road(Transportation)" tag instead.

Tables 4.1, 4.2, 4.3, 4.4 consist of the early stage data collected during December 2015 by four different participants. Two of the participants are faculty members, one of them is a master student and the last one is an undergraduate student.

Label	True Positive	False Positive
Home	0,890	0,006
On the Road (Transportation)	0,919	0,065
Work/School	0,973	0,025
Canteen/Restaurant/Cafe/Bar	0,864	0,009
Other	0,850	0,038
Boat	0,000	0,000
Yellow Minibus	0,664	0,000
Walk	0,000	0,000
Weighted Average	0,885	0,029

Table 4.1: December Data Classification with DT Algorithm Male Faculty Member

Label	True Positive	False Positive
On the Road (Transportation)	0,988	0,053
Home	0,940	0,000
Canteen/Restaurant/Cafe/Bar	0,901	0,000
Work/School	0,996	0,008
Weighted Average	0,972	0,032

Table 4.2: December Data Classification with DT Algorithm Female Faculty Member

Label	True Positive	False Positive
Sit	1,000	0,011
On the Road (Transportation)	0,950	0,080
[GSU]Canteen/Garden	1,000	0,044
Work/School	0,625	0,000
Outdoor(Park, etc.)	0,000	0,000
Weighted Average	0,906	0,041

Table 4.3: December Data Classification with DT Algorithm undergraduate student

Label	True Positive	False Positive
Home	0,948	0,015
On the Road (Transportation)	0,840	0,029
Mall/Shop	0,932	0,008
Canteen/Restaurant/Cafe/Bar	0,911	0,000
Work/School)	0,952	0,043
Gym	0,937	0,007
Other	0,000	0,000
Station	1,000	0,008
Weighted Average	0,917	0,027

Table 4.4: December Data Classification with DT Algorithm Master Student

All the data is classified with using only Time Features. As you can see even though “Home” label is one of the main labels of every participant’s daily route ; the undergraduate student never used it. The label "Sit" is probably used instead of our "Home" label.In order to overcome this problem in the second stage of our data collection a predefined label list was given to all the participants.

In the second stage of our data collection an upgrade is made on the ARService application to give user a predefined label list to choose. This pop-up menu appears in every fifteen minutes and asks the user for his/her location. Similarly, for the manual location markup a predefined list of locations was given but the user can define and new locations.

In the second stage, two participants, one female faulty member and one master student, collected data using two smart phones for one month. The results with J48 Decision Tree classification algorithm are given in Tables 4.5 and 4.6. Both tables present the results classified with Decision Tree J48 algorithm with Time Feature only.The devices are Samsung Galaxy S5 and Samsung Galaxy S3 Mini smart phones. Two participants used these smart phones at the same time. Even though the same participant used these two smart phones at the same time ; collected data has shown differences.

Label	True Positive	False Positive
Canteen/Restaurant/Cafe/Bar	0,361	0,043
On the Road (Transportation)	0,701	0,149
Work/School	0,354	0,006
Outdoor (Park, etc.)	0,547	0,005
Home	0,772	0,055
Other	0,774	0,081
[GSU]Classroom/Lab	0,853	0,038
Mall/Shop	0,743	0,024
Cinema	0,398	0,007
Gym	0,000	0,000
Weighted Average	0,667	0,076

Table 4.5: Galaxy S5 Data, Female Faculty Member, J48

Label	True Positive	False Positive
On the Road (Transportation)	0,988	0,053
Home	0,940	0,000
Canteen/Restaurant/Cafe/Bar	0,901	0,000
Work/School	0,996	0,008
Weighted Average	0,972	0,032

Table 4.6: Galaxy S3 Mini Data, Female Faculty Member, J48

Samsung Galaxy S5 data have more labels. However Samsung Galaxy S3 Mini data has a better TP rate then Samsung Galaxy S5 data. While "Home" label has 0,772 percentage of true positive prediction with Samsung Galaxy S5 data; "Home" label has 0.940 true positive prediction rate with Samsung Galaxy S3 Mini data. The difference between the same tags might be because of the participant's tagging habit. Since Samsung Galaxy S5 smart phone is the participants main device, the participant can easily tag more notifications. While Samsung Galaxy S3 Mini device is used less frequently, the tagging was more proficient and the number of classes was lower.

The second participant, whom used Samsung Galaxy S5 and Samsung Galaxy S3 Mini smart phones at the same time had similar results, shown in Tables 4.8 and 4.7. This participant also used Samsung Galaxy S5 device as the main phone, so notifications were tagged more often by that. Again the data labels differ from each other. The main phone Galaxy S5 has an additional label "Canteen/Restautant/Cafe/Bar".

Label	True Positive	False Positive
[GSU]Classroom/Lab	0.268	0,000
Canteen/Restautant/Cafe/Bar	1,000	0,039
Outdoor (Park, etc.)	0,932	0,238
On the Road (Transportation)	0,640	0,138
Mall/Shop	0,000	0,000
Home	0,871	0,000
Weighted Average	0,716	0,132

Table 4.7: S5 Data

Label	True Positive	False Positive
[GSU]Classroom/Lab	1,000	0,000
On the Road (Transportation)	1,000	0,563
Outdoor(Park, etc.)	0,000	0,000
Mall/Shop	0,000	0,000
Home	1,000	0,000
Weighted Average	0,714	0,304

Table 4.8: S3 Mini Data

The result obtained from our preliminary research has been published as (Çelik and İncel, 2016) in the 2016, 24th Signal Processing and Communication Application Conference (SIU).

4.2 Data Visualization

After the preliminary data collection and correcting the problems encountered in this phase, we collected data from 17 participants. A summary about the participants is presented in ??.

As you can see in the ?? in the last stage of our data collection we had 20 participant candidates. However, after preprocessing the collected data; we discovered that some of the data was not suitable to include our research. Some smart phones had problem with server communication and refuse to send data or corrupt it, some candidates collected below our threshold instances, so it was unnecessary to include them.

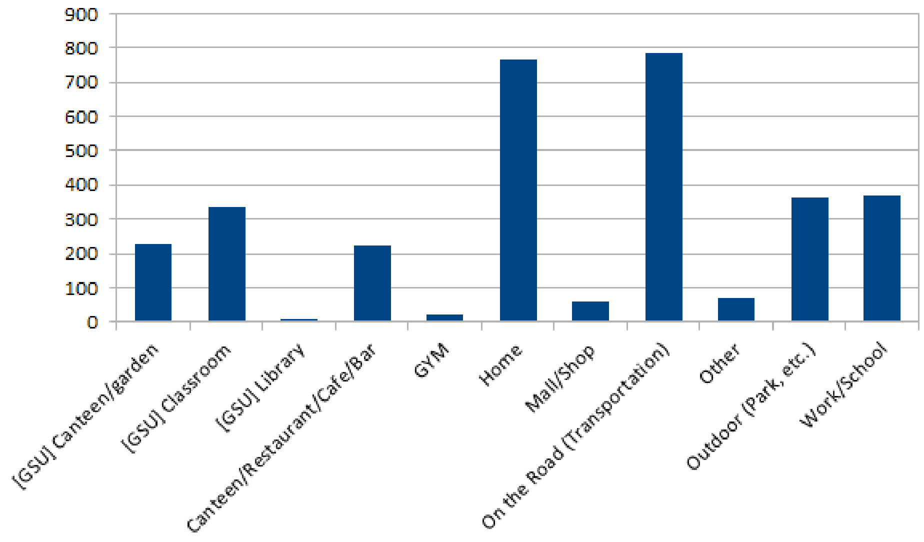


Figure 4.1: Number of Labels for Place Categories

No	Age	Gender	Class	Occupation	Smart Phone Model
1	21	M	1	Student	Samsung S3 Mini
2	20	M	1	Student	LG G2
3	28	F	Master Student	Full Timer	Samsung S5
4	35	F	Faculty Member	Full Timer	Samsung S5
5	21	F	3	Part Timer/Student	Samsung S4
6	21	F	1	Student	Samsung ST7580
7	20	M	1	Student	Meizu MX4
8	42	M	Faculty Member	Full Timer	Samsung S3 Neo
9	23	F	4	Student	Samsung J2
10	21	M	3	Student	General Mobile 4G
11	23	F	4	Part Timer/Student	Samsung J7
12	19	M	1	Student	Samsung S3 Neo
13	10	M	1	Student	Samsung S5
14	22	K	4	Student	Samsung Note 2
15	19	M	1	Student	Samsung S4 Mini
16	18	M	Prep.	Student	Samsung S3
17	22	M	4	Student	Samsung S4
18	19	M	1	Student	Samsung S3 Mini
19	22	M	4	Student	Samsung S3 Mini
20	23	M	4	Student	Samsung Note 2

Table 4.9: Participant Table

Before processing the collected data from 17 participants, we visually explored the relationship between the place tags and the feature types extracted from the raw data. Although we apply feature selection methods to the raw dataset in Section 4.4, our aim was to decide on the set of features by analyzing the raw data initially.

In Figure 4.1, a summary of the data collected from 17 people is presented. In this figure the total number of tags for each place category is presented. Each tag is given to a place in the total number of units entered.

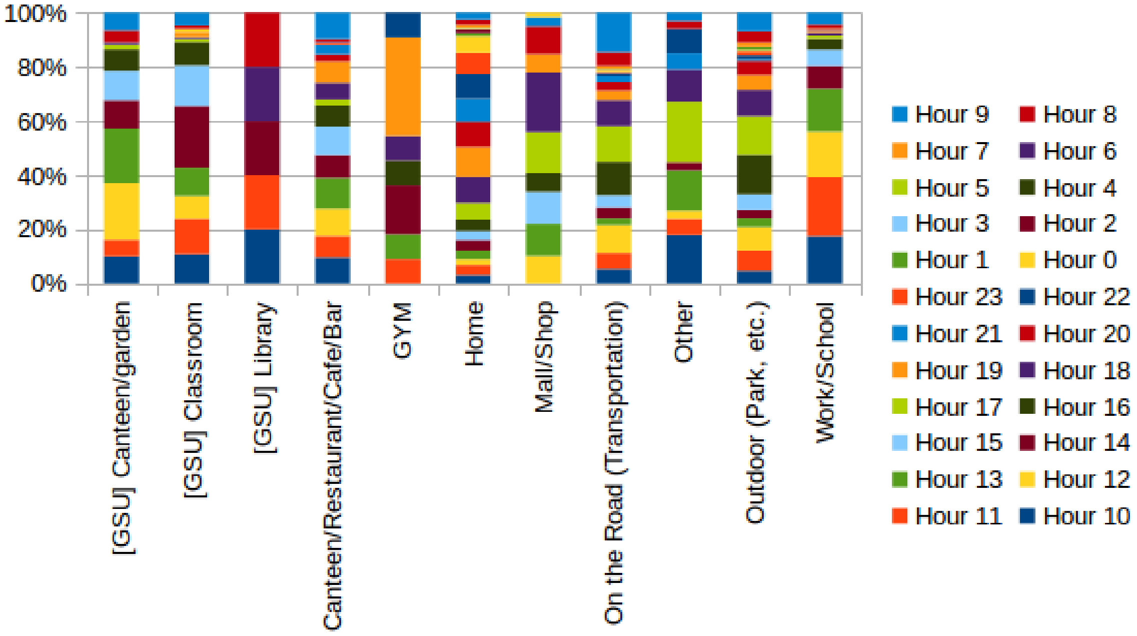


Figure 4.2: Places and their visit hours during a day

Time Features : In Figures 4.2, 4.3 and 4.4 we present the number of labels for time features : hour, time of the day, week/weekend, respectively. In Figure 4.4, we observe that place types such as work/school, GSU Canteen, GSU Classroon and Library are not labeled at the weekends, while the Home category is the mostly tagged place at the weekends, as expected. In Figure 4.3, it is seen that while “Home” is the mostly tagged place in the evenings, work/school and other classes related to GSU campus are tagged mostly in the mornings and afternoons. The places tagged at night are Home and on the road/transportation while other place types were labelled at night very seldom or not at all.

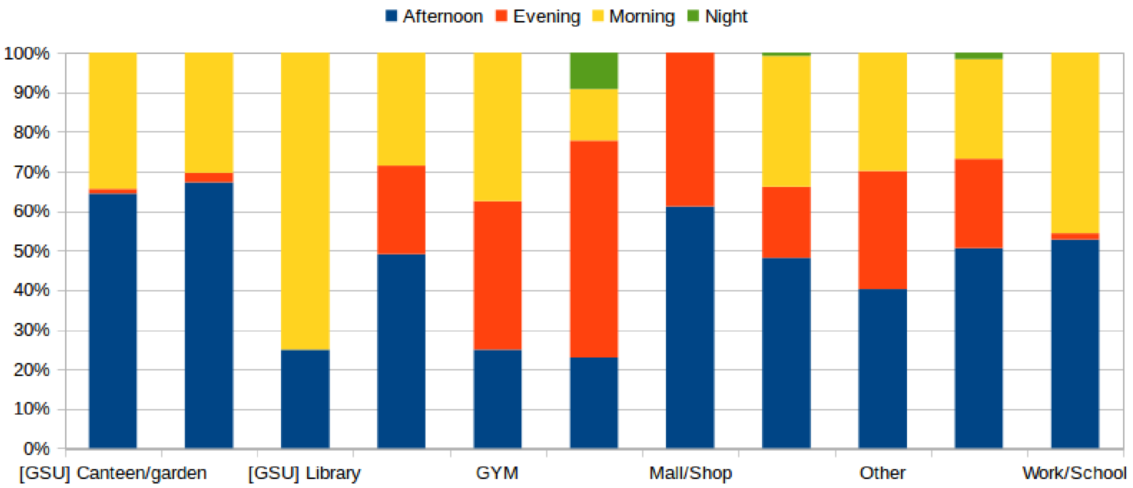


Figure 4.3: Places and their visit times during a day

Hour Feature is one of the most efficient features we have. For almost every place label, we can make prediction with using Hour Feature. As in the 4.2 we can see the impact of it. Hour 17 (time between 17 :00 and 17 :59) is mostly points to "On the Road(Transportation)" label. While people tend to at the GYM at the Hour 21 (21 :00 - 21 :59) period, we can say Hour 13 (13 :00 - 13 :59) period mostly used as lunch break based on [GSU]Canteen/garden and Canteen/Restaurant/Cafe/Bar label behaviours. Some Hour periods have almost no tagging. These are Hour 2, Hour 3 and Hour 4. These hours are (between 02 :00 - 04 :59) mostly spent in sleeping. So the absence of the tagging was expected.

Communication Information Features : In Figure 4.5, we present the average and maximum number of BSID' s recorded at different place categories. Similarly, in Figure 4.6, we present average and maximum number of SSID' s at different places. When a participant is at a particular place, the ARService application may have recorded signals from a specific base station. Instead of using the total number of records, we use the unique number of base stations and access points at a particular place. When the participants are changing location, such as on the road/transportation, or walking outdoors (Outdoor (Park, etc.)), the average and maximum number of SSID' s and BSID' s are higher.

Phone Features : Another feature set that we analyze is the phone features. This set includes the number of times the phone screen is turned on and turned off, number of times the headset is plugged in, number of running applications and battery levels.

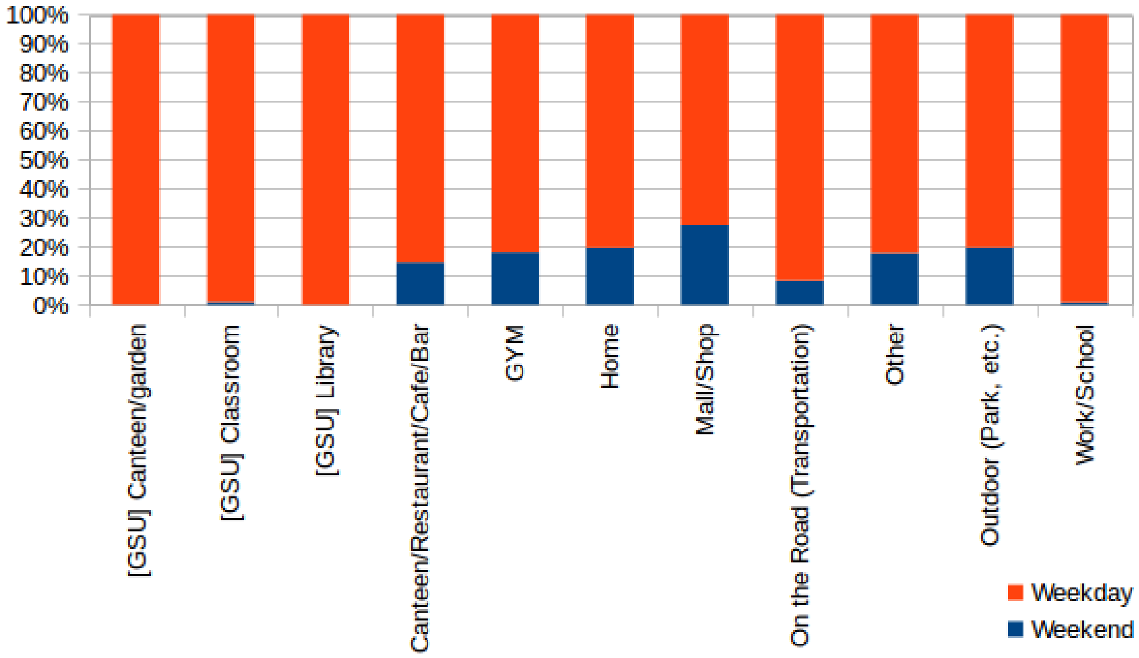


Figure 4.4: Places and Week/Weekend Relationship

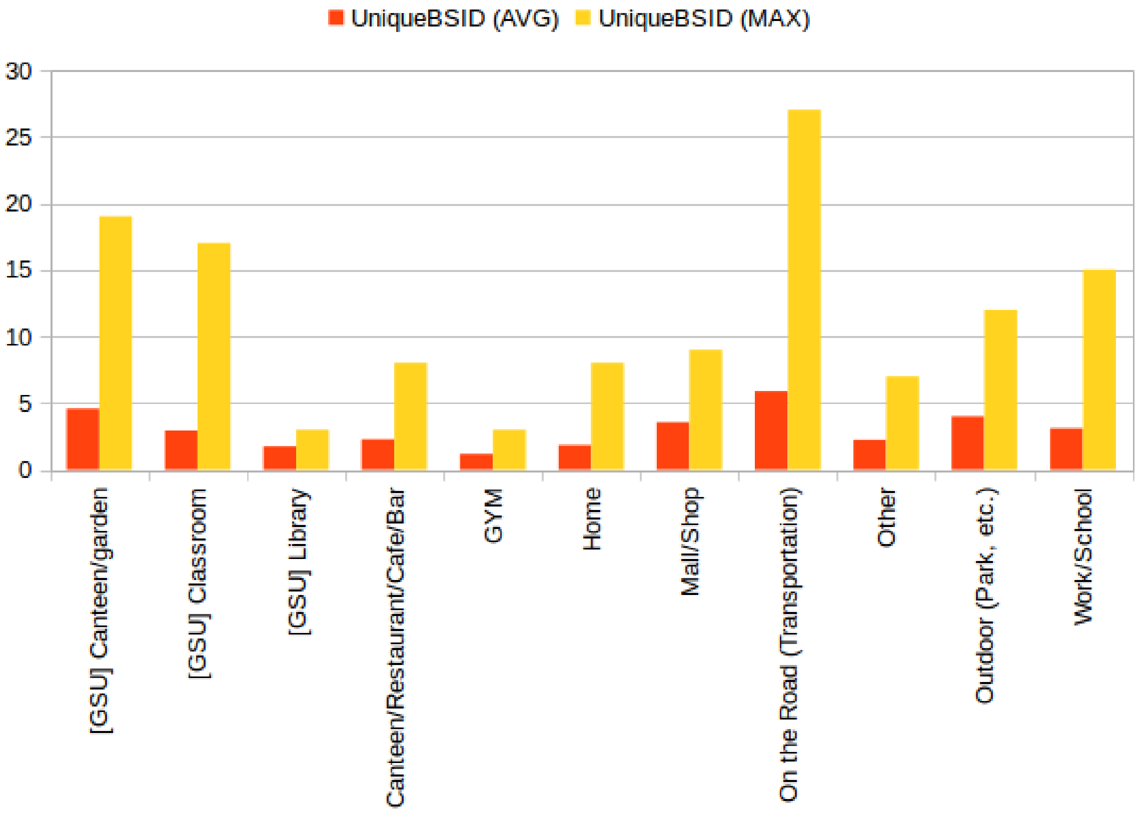


Figure 4.5: Number of unique BSID' s from which a signal is received at a place

In Figure 4.7, we present the place and phone usage relationship. In Figure 4.8 we present the battery levels and places, and in Figure 4.9, we present the number of

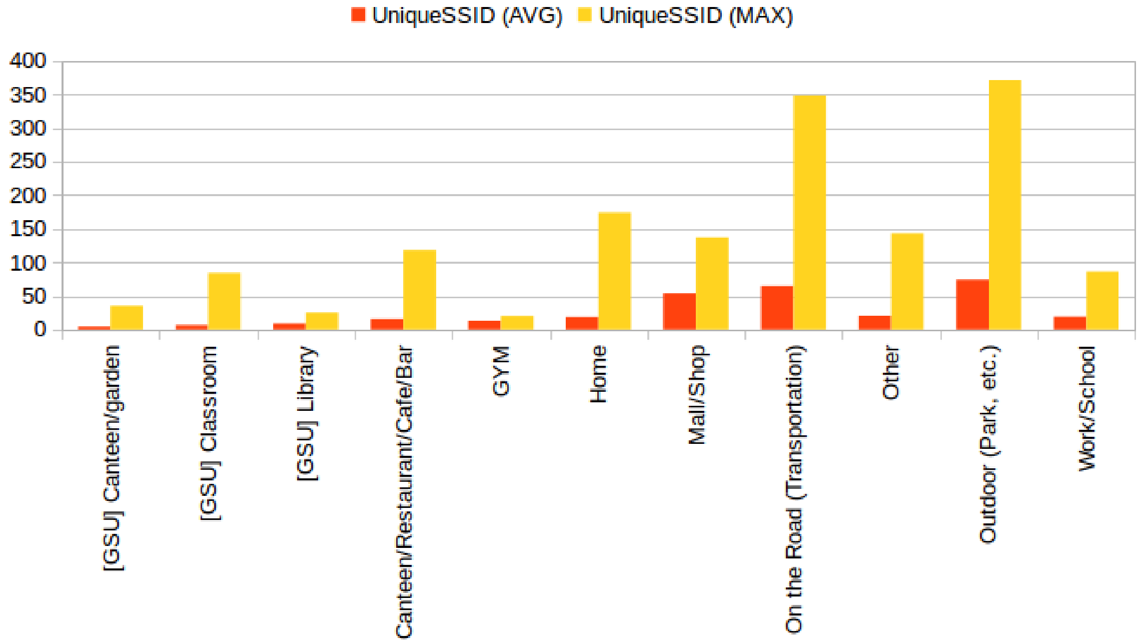


Figure 4.6: Number of unique SSID' s from which a signal is received at a place

running places versus place categories.

For battery Levels, instead of using the exact levels, we grouped the levels as : 20-40%, 40-60%, 60-80% and 80-100%. This feature helps us to understand the participant's phone usage behaviour.

Activity Features : One of the questions that we try to answer in this thesis is that whether there is a relationship between the places and the activities taking place in such places. Another feature set that we utilized is the activities of the participants identified by the ARService application in real-time.

In Figure 4.10, we visualize the relationship between the places and the state of

users, i.e., whether they are still or mobile (such as walking, transportation). At places, such as Home, canteen/restaurant/cafe/bar, classroom, library, participants are mostly stationary. On the other hand, at places, such as outdoor, mall/shop, on the road, participants are mobile, as expected. One unexpected result is observed at gym. While we expect the participants to be mobile here, they are observed to be stationary. When we questioned the users about this, they replied that the phone was left in a drawer or on a flat surface while exercising. If a smart watch was used while collecting data, this situation could be eliminated.

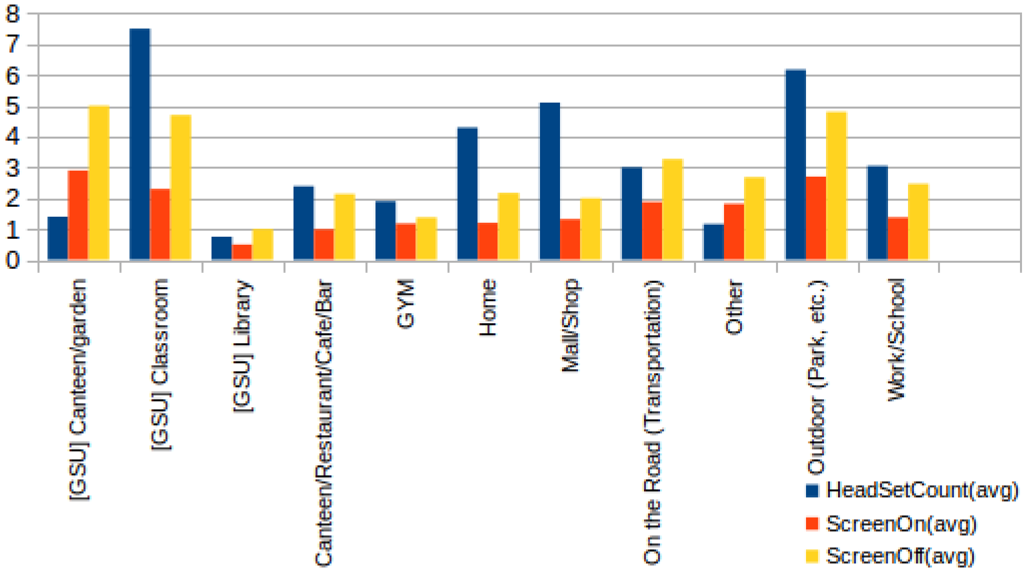


Figure 4.7: Phone Usage at Different Places : screen and headset states

In Figure 4.11, we present the relationship between the places and activities taking places in the places. Y-axis is given in %. The activities were not marked by users but ARService automatically predicts the activities, with around 80% accuracy. Here, we observe that in mall/shop and outdoor, walking activity is mostly recognized, while at home sitting and standing activities were mostly seen. For on the road (transportation) class, percentage of transportation activity is similar to other activities. The reason is that, ARService classifies the activity as sitting and standing while the vehicle is stationary, either in a traffic jam or at traffic lights.

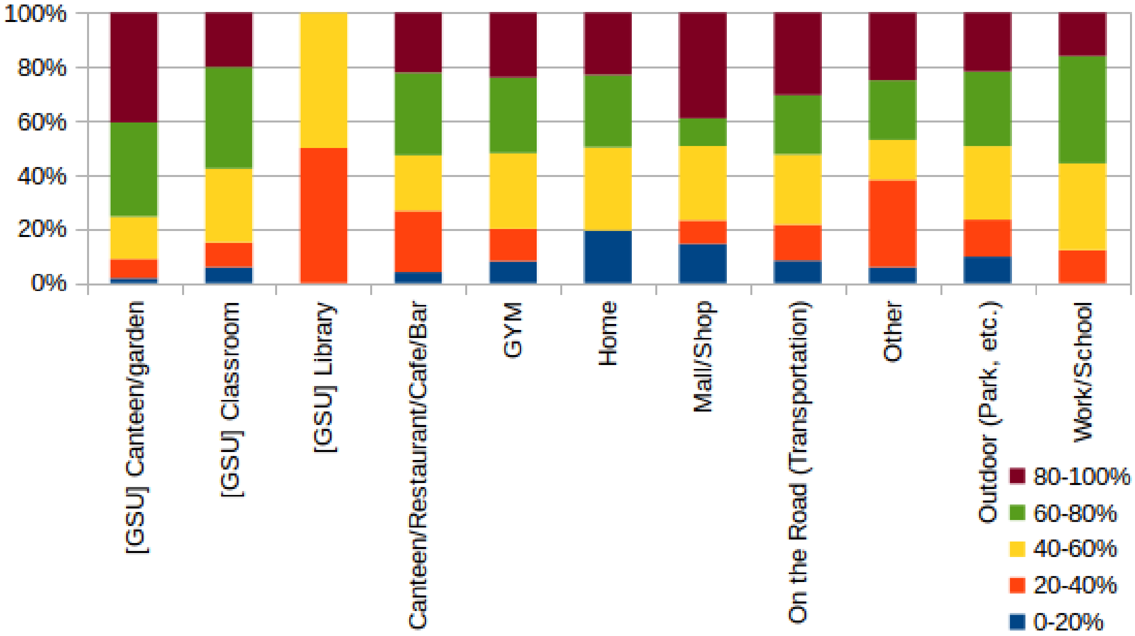


Figure 4.8: Battery Level Distribution

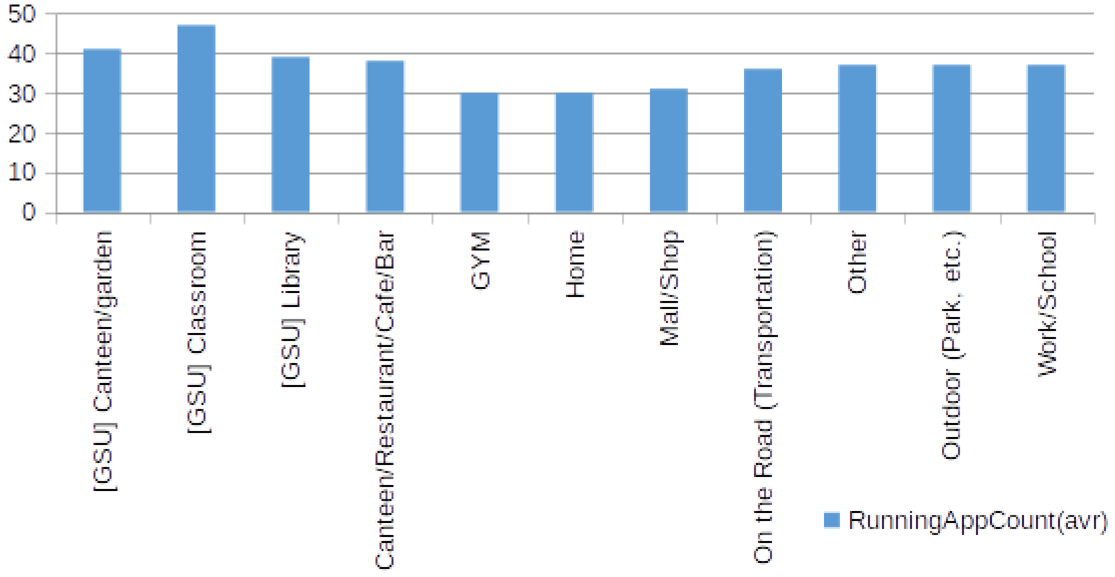


Figure 4.9: Average Number of Running Applications

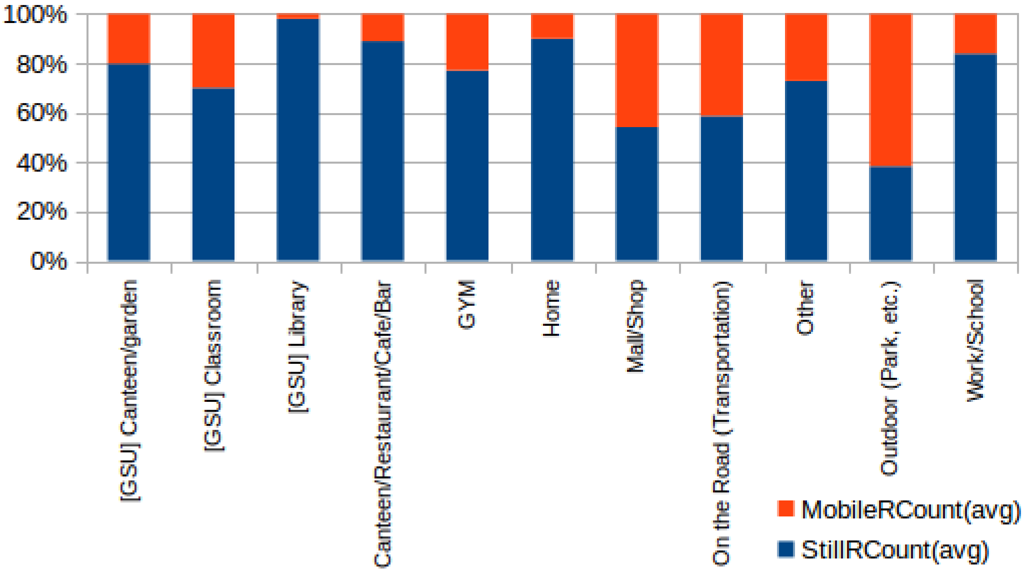


Figure 4.10: Whether the participant is mobile or stationary at a specific place (given in %)

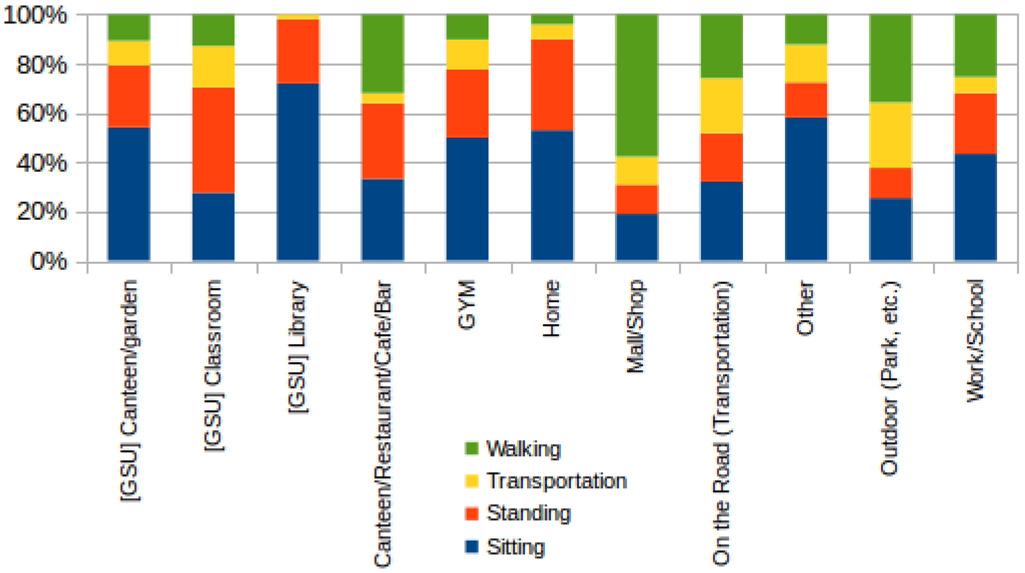


Figure 4.11: Place and Activity Relationship

4.3 Impact of Different Feature Combinations

In this section, we analyze the classification performance with different feature sets using four different classification algorithms, namely Naive Bayes, KNN, decision tree and random forest.

In Figure 4.12, we present the results with different feature combinations. The letters, i.e., the short forms, used in the figure are listed as follows :

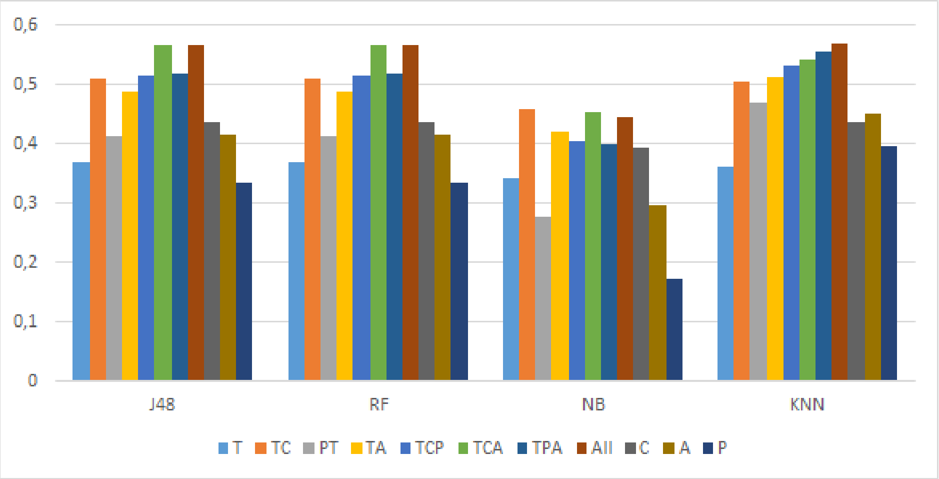


Figure 4.12: Feature Combinations with 14 Participant Data

- T : Time Feature only. Consists of Day, Hour, Week/Weekend and ToD (Time of Day) features.
- C : Communication Information Feature only. Contains UniqueBSID and UniqueSSID features.
- A : Activity Features only. Contains StillCount, MobileCount, SittingCount, StandingCount, TransportationCount, WalkingCount, RunningCount, StairsCount features.
- P : Phone Features only. They are RunningAppCount, HeadSetCount, ScreenON, ScreenOFF, and BatteryThreshold features.
- TC :Combination of Time and Communication Information features.
- TP : Combination of Time and Phone Features.
- TA :Combination of Time and Activity Features.
- TCP :Combination of Time, Communication Information and Phone Features.
- TCA :Combination of Time, Communication Information and Activity Features.

- TPA :Combination of Time, Phone and Activity Features.
- All : Combination of the all features we have, Time, Communication Information, Phone and Activity Features.

In Figure 4.12, the y-axis is given in terms of TP (True positive) rate. The results are presented with Decision Tree(J48), Random Forest, Naive Bayes and K-Nearest Neighbour classification algorithms. We observe that, Phone Features alone have the lowest success rate for the all four classification algorithms. However, when they are combined with Time Features, success rates increase at least by 0.1. We observe that some feature combinations return better prediction rates than others. Features alone have lower recognition rates then binary or triple feature combinations.

Time Feature returns almost the same rates with all the four classification algorithms. Same observation can be mentioned for CommunicationInformation Features as well.

When Phone Features are combined with Naive Bayes they return the lowest true positive rate ; while combining with KNN, it increases to higher rates.

When we combine time and other three features one by one ; all the binary combinations give better results. Among all three binary feature combinations ; TC exhibit the highest rates among all the four classification algorithms.

Time Features (Day, Hour, ToD and Week.Weekend) let us to categorize our participants' routine behaviours. A student, or a faculty member tends to have classes in the morning- afternoon- evening times, mostly at the weekdays. In the light of these behaviours ; our algorithm has a higher possibility to predict their whereabouts when combining with other features.

For CommunicationInformation features UniqueSSID and UniqueBSID numbers tend to change as we move around. The number of access points and base stations change when a participant changes the location. With this information, our algorithm can have higher prediction rates.

For the triple combinations of TCP, TCA, TPA ; the TCA has the highest rates. TC alone has a prediction rate around 50% success rate with the all four classification algorithms.

Activity Features are telling us if the participant is on foot, sitting, walking, etc. When a participant is walking, or in transport mode this knowledge combined with time variables (Monday at 8, in the morning) and CommunicationInformation (increase and decrease of the UniqueSSID and UniqueBSID numbers) the classification algorithms exhibit better results.

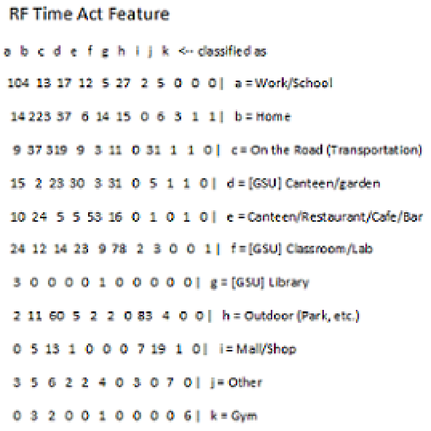
When all features are combined, they exhibit the most highest rate as we predicted. However TCA gives almost the same recognition rates. In Section 4.4, we apply feature selection algorithms to our data and this reveals that time, communication and activity features are the mostly selected features compared to phone features.

Decision tree algorithms J48 and Random Forest presented almost the same results, while K-Nearest Neighbour algorithm followed them with good results. But the Naive Bayes algorithm has the lowest results.

As an example, the confusion matrices for the random forest classifier is given in Table 4.10. On the Road(Transportation) and Outdoor(Park, etc) labels tend to be confused with each other. Since both of them are placed outside the change of Activity, Communication Information and Phone Features lead to this confusion. While a bus stop is located at the Outdoor (Park, etc), the bus itself counts as On the Road(Transportation). Home, Work/School and On the Road(Transportation) labels also get mixed. For all the feature combinations, the number of confused prediction times change. By using only Time Features, the algorithm mixed Work/School with Home 13 times and mixed with On the Road(Transportation) 17 times. These numbers change as ; mixed Work/-School with Home 23 times. And mixed with On the Road(Transportation) 16 times by using Time and CommunicationInformation Features. By using Time and Activity Feature combination we got mixed Work/School with Home 15 times and mixed with On the Road(Transportation) 17 times. No drastic changes are observed here. When we use all features together the numbers are 10 times for Home and 29 times for On the Road(Transportation). On the other hand, Home label is mixed with Work/School 11 times and with On the Road(Transportation) 14 times when we combine all features. For the On the Road(Transportation), 6 times Work/School is chosen, while 24 times Home is chosen instead of it. Surprisingly one of the other Home place confusion happened between Canteen/Restaurant/Cafe/Bar. By using Time and Activity Features combination Canteen/Restaurant/Cafe/Bar is mixed with Home 24 times. Time Feature also got confused with Home 24 times again. With Time and Phone Features combination the mix decreases to 17 times.

One of the most repeated confusion happened between [GSU]Library and [GSU]Canteen/garden. It may happen because the Galatasaray University Library is located inside the garden. Canteen and library are within 1 minute walking distance. Using all feature combinations together decreased these confusions.

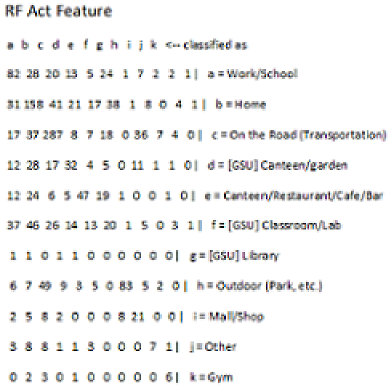
Time



```
RF Time Act Feature

a b c d e f g h i j k <-- classified as
104 13 17 12 5 27 2 5 0 0 0 | a = Work/School
14 223 37 6 14 15 0 6 3 1 1 | b = Home
9 37 319 9 3 11 0 31 1 1 0 | c = On the Road (Transportation)
15 2 23 30 3 31 0 5 1 1 0 | d = [GSU] Canteen/garden
10 24 5 5 53 16 0 1 0 1 0 | e = Canteen/Restaurant/Cafe/Bar
24 12 14 23 9 78 2 3 0 0 1 | f = [GSU] Classroom/Lab
3 0 0 0 0 1 0 0 0 0 0 | g = [GSU] Library
2 11 60 5 2 2 0 83 4 0 0 | h = Outdoor (Park, etc.)
0 5 13 1 0 0 0 7 19 1 0 | i = Mall/Shop
3 5 6 2 2 4 0 3 0 7 0 | j = Other
0 3 2 0 0 1 0 0 0 0 6 | k = Gym
```

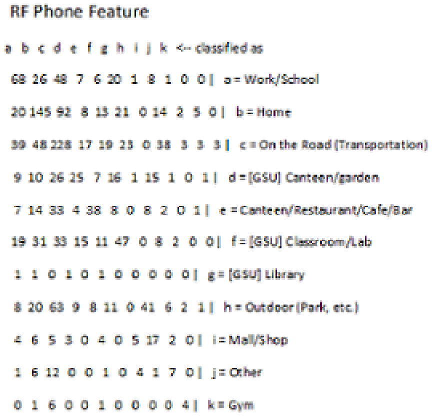
Activity



```
RF Act Feature

a b c d e f g h i j k <-- classified as
82 28 20 13 5 24 1 7 2 2 1 | a = Work/School
31 158 41 21 17 38 1 8 0 4 1 | b = Home
17 37 287 8 7 18 0 36 7 4 0 | c = On the Road (Transportation)
12 28 17 32 4 5 0 11 1 1 0 | d = [GSU] Canteen/garden
12 24 6 5 47 19 1 0 0 1 0 | e = Canteen/Restaurant/Cafe/Bar
37 46 26 14 13 20 1 5 0 3 1 | f = [GSU] Classroom/Lab
1 1 0 1 1 0 0 0 0 0 0 | g = [GSU] Library
6 7 49 9 3 5 0 83 5 2 0 | h = Outdoor (Park, etc.)
2 5 8 2 0 0 0 8 21 0 0 | i = Mall/Shop
3 8 8 1 1 3 0 0 0 7 1 | j = Other
0 2 3 0 1 0 0 0 0 0 6 | k = Gym
```

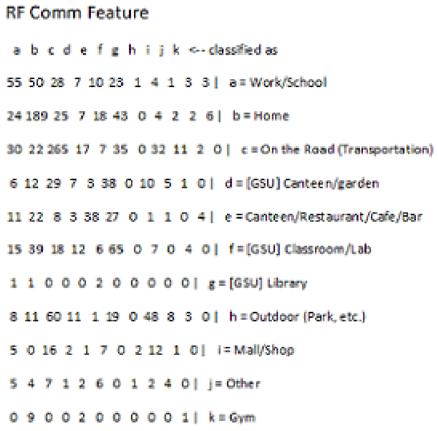
Phone



```
RF Phone Feature

a b c d e f g h i j k <-- classified as
68 26 48 7 6 20 1 8 1 0 0 | a = Work/School
20 145 92 8 13 21 0 14 2 5 0 | b = Home
39 48 228 17 19 23 0 38 3 3 3 | c = On the Road (Transportation)
9 10 26 25 7 16 1 15 1 0 1 | d = [GSU] Canteen/garden
7 14 33 4 38 8 0 8 2 0 1 | e = Canteen/Restaurant/Cafe/Bar
19 31 33 15 11 47 0 8 2 0 0 | f = [GSU] Classroom/Lab
1 1 0 1 0 1 0 0 0 0 0 | g = [GSU] Library
8 20 63 9 8 11 0 41 6 2 1 | h = Outdoor (Park, etc.)
4 6 5 3 0 4 0 5 17 2 0 | i = Mall/Shop
1 6 12 0 0 1 0 4 1 7 0 | j = Other
0 1 6 0 0 1 0 0 0 0 4 | k = Gym
```

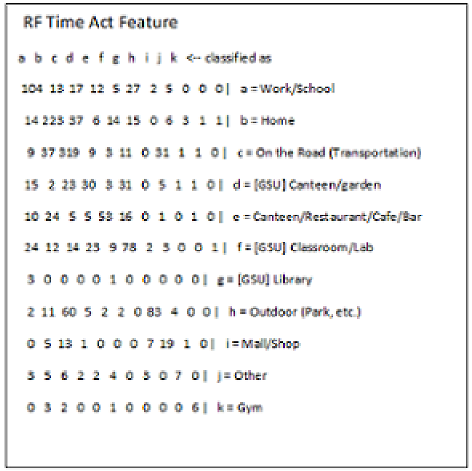
Communication



```
RF Comm Feature

a b c d e f g h i j k <-- classified as
55 50 28 7 10 23 1 4 1 3 3 | a = Work/School
24 189 25 7 18 43 0 4 2 2 6 | b = Home
30 22 265 17 7 35 0 32 11 2 0 | c = On the Road (Transportation)
6 12 29 7 3 38 0 10 5 1 0 | d = [GSU] Canteen/garden
11 22 8 3 38 27 0 1 1 0 4 | e = Canteen/Restaurant/Cafe/Bar
15 39 18 12 6 65 0 7 0 4 0 | f = [GSU] Classroom/Lab
1 1 0 0 0 2 0 0 0 0 0 | g = [GSU] Library
8 11 60 11 1 19 0 48 8 3 0 | h = Outdoor (Park, etc.)
5 0 16 2 1 7 0 2 12 1 0 | i = Mall/Shop
5 4 7 1 2 6 0 1 2 4 0 | j = Other
0 9 0 0 2 0 0 0 0 0 1 | k = Gym
```

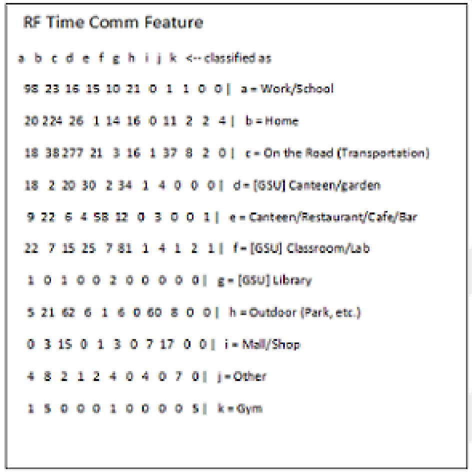
Time and Activity



```
RF Time Act Feature

a b c d e f g h i j k <-- classified as
104 13 17 12 5 27 2 5 0 0 0 | a = Work/School
14 223 37 6 14 15 0 6 3 1 1 | b = Home
9 37 319 9 3 11 0 31 1 1 0 | c = On the Road (Transportation)
15 2 23 30 3 31 0 5 1 1 0 | d = [GSU] Canteen/garden
10 24 5 5 53 16 0 1 0 1 0 | e = Canteen/Restaurant/Cafe/Bar
24 12 14 23 9 78 2 3 0 0 1 | f = [GSU] Classroom/Lab
3 0 0 0 0 1 0 0 0 0 0 | g = [GSU] Library
2 11 60 5 2 2 0 83 4 0 0 | h = Outdoor (Park, etc.)
0 5 13 1 0 0 0 7 19 1 0 | i = Mall/Shop
3 5 6 2 2 4 0 3 0 7 0 | j = Other
0 3 2 0 0 1 0 0 0 0 6 | k = Gym
```

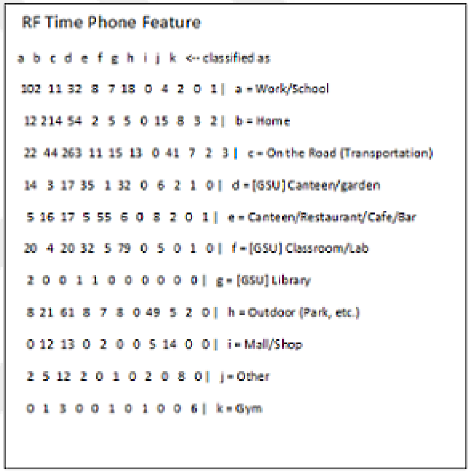
Time and Communication



```
RF Time Comm Feature

a b c d e f g h i j k <-- classified as
98 23 16 15 10 21 0 1 1 0 0 | a = Work/School
20 224 26 1 14 16 0 11 2 2 4 | b = Home
18 38 277 21 3 16 1 37 8 2 0 | c = On the Road (Transportation)
18 2 20 30 2 34 1 4 0 0 0 | d = [GSU] Canteen/garden
9 22 6 4 58 12 0 3 0 0 1 | e = Canteen/Restaurant/Cafe/Bar
22 7 15 25 7 81 1 4 1 2 1 | f = [GSU] Classroom/Lab
1 0 1 0 0 2 0 0 0 0 0 | g = [GSU] Library
5 21 62 6 1 6 0 60 8 0 0 | h = Outdoor (Park, etc.)
0 3 15 0 1 3 0 7 17 0 0 | i = Mall/Shop
4 8 2 1 2 4 0 4 0 7 0 | j = Other
1 5 0 0 0 1 0 0 0 0 5 | k = Gym
```

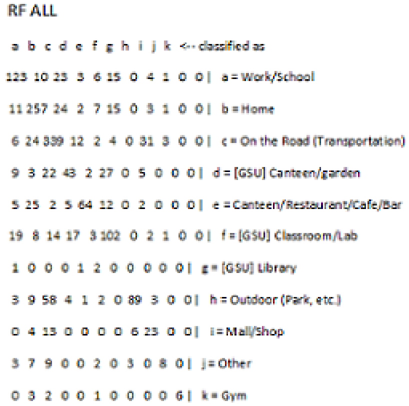
Time and Phone



```
RF Time Phone Feature

a b c d e f g h i j k <-- classified as
102 11 32 8 7 18 0 4 2 0 1 | a = Work/School
12 214 54 2 5 5 0 15 8 3 2 | b = Home
22 44 263 11 15 13 0 41 7 2 3 | c = On the Road (Transportation)
14 3 17 35 1 32 0 6 2 1 0 | d = [GSU]Canteen/garden
5 16 17 5 55 6 0 8 2 0 1 | e = Canteen/Restaurant/Cafe/Bar
20 4 20 32 5 79 0 5 0 1 0 | f = [GSU] Classroom/Lab
2 0 0 1 1 0 0 0 0 0 0 | g = [GSU] Library
8 21 61 8 7 8 0 49 5 2 0 | h = Outdoor (Park, etc.)
0 12 13 0 2 0 0 5 14 0 0 | i = Mall/Shop
2 5 12 2 0 1 0 2 0 8 0 | j = Other
0 1 3 0 0 1 0 1 0 0 6 | k = Gym
```

All



```
RF ALL

a b c d e f g h i j k <-- classified as
123 10 23 3 6 15 0 4 1 0 0 | a = Work/School
11 257 24 2 7 15 0 3 1 0 0 | b = Home
6 24 339 12 2 4 0 31 3 0 0 | c = On the Road (Transportation)
9 3 22 43 2 27 0 5 0 0 0 | d = [GSU] Canteen/garden
5 25 2 5 64 12 0 2 0 0 0 | e = Canteen/Restaurant/Cafe/Bar
19 8 14 17 3 102 0 2 1 0 0 | f = [GSU] Classroom/Lab
1 0 0 0 1 2 0 0 0 0 0 | g = [GSU] Library
3 9 58 4 1 2 0 89 3 0 0 | h = Outdoor (Park, etc.)
0 4 13 0 0 0 0 6 23 0 0 | i = Mall/Shop
3 7 9 0 0 2 0 3 0 8 0 | j = Other
0 3 2 0 0 1 0 0 0 0 6 | k = Gym
```

Table 4.10: Random Forest Label Confusion Matrix According to Feature Combination

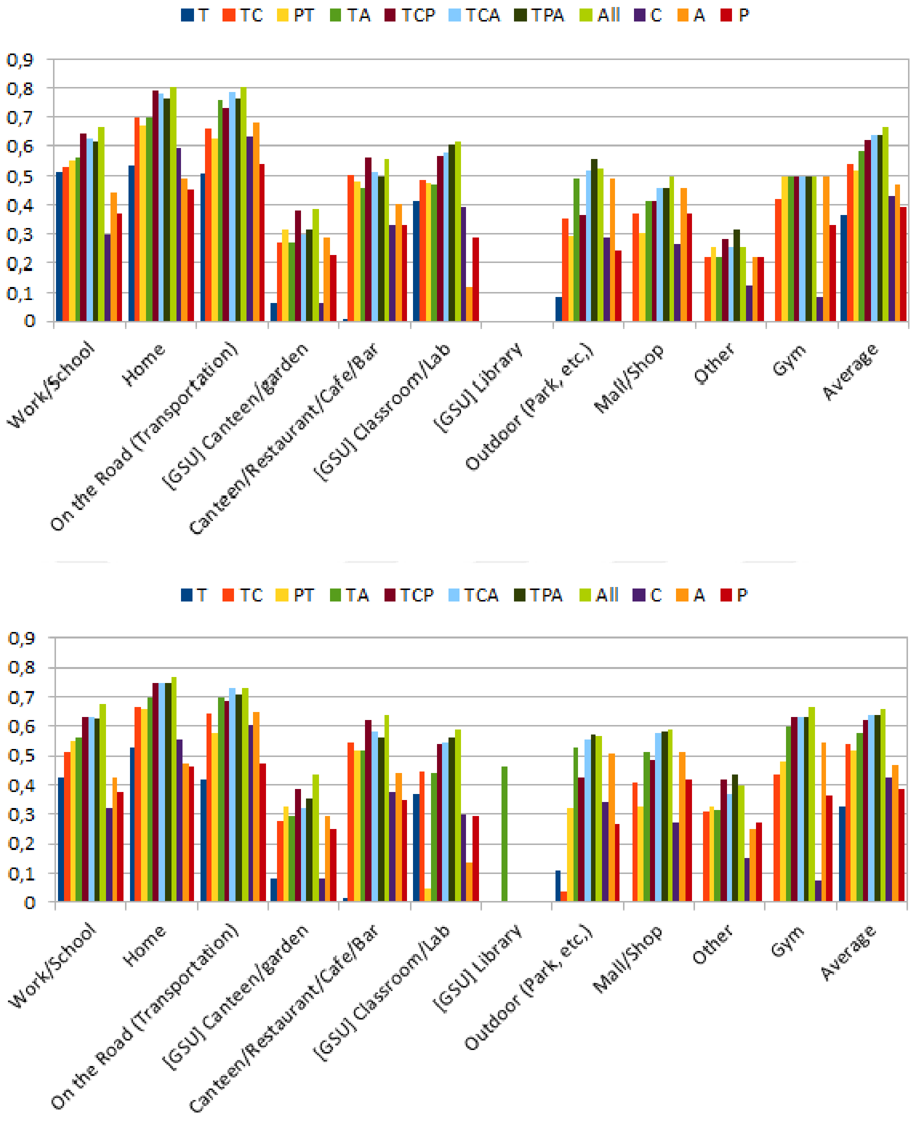


Figure 4.13: True Positive and F Measure Results for Random Forest Classification

In Figure 4.13, we provide class-specific results. True Positive and F Measure results using Random Forest classification algorithms are obtained. [GSU] Library label was never predicted correctly since the number of instances in the whole data was very low (Figure 4.1). While Home label was predicted with good results, we can not say the same thing for [GSU]Canteen/garden. Especially with using Phone and Time Features combinations, higher false positive rates are obtained. This is due to fact that, [GSU]Canteen/garden is confused with Work/School (for all participants GSU Campus was labelled as Work/School) and with GSU Classroom/Lab classes as we see in Table 4.10. The "Other" label contains not periodically visited places. Since periodic movements could be predicted using Time features, some irregularities could be mixed with each other. When a participant leaves home 30 minutes before than usual hour, prediction might confuse it. For the Other label this problem does not exist. It is already an irregular behaviour.

Both True positive and f measure has high level results for Home, On the Road(Transportation) labels. Work/School label follows them closely. These three labels alone have huge impact in our daily lives. All of our participants are either students or have a full time job as faculty members. This is reflected in the results.

4.4 Impact of Feature Selection

In Section 4.3, we observed that different feature sets exhibit different performance results. In this section, we apply different feature selection algorithms to explore which features, instead feature sets, are more efficient in classifying the places.

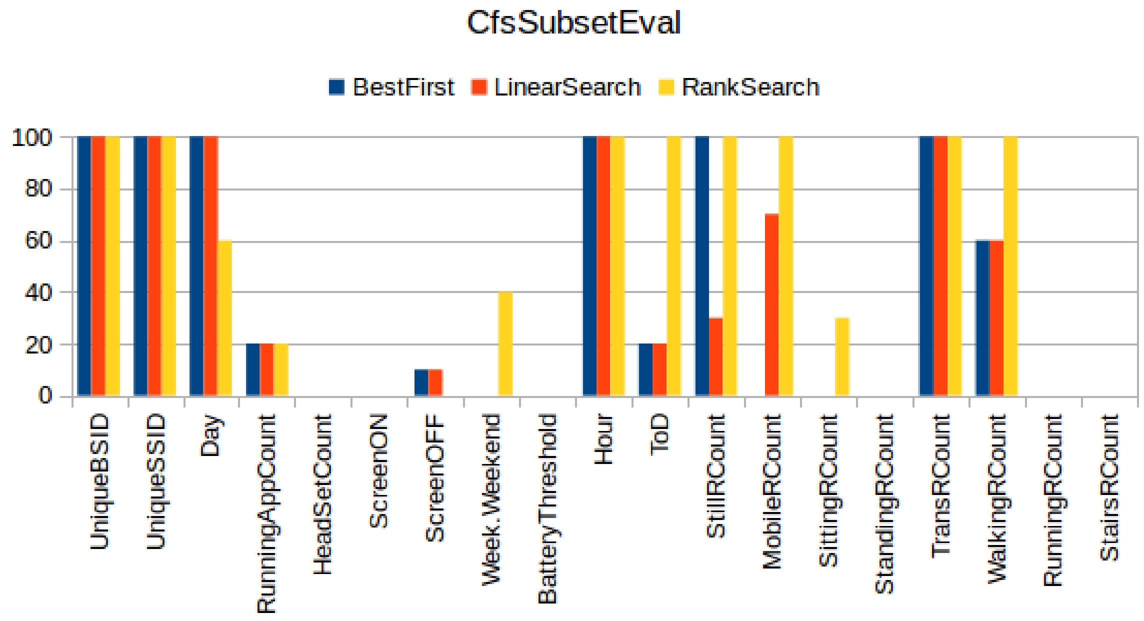


Figure 4.14: CfsSubsetEval Results for All Participants

We use CfsSubsetEval and ClassifierSubsetEval Attribute Selection algorithms (explained in Section 3.1.3) are used with J48, Random Forest, Naive bayes and K Nearest Neighbour algorithms. As the search methods, we use BestFirst, LinearForwardSearch and RankSearch methods explained in Section 3.1.3. Results are given for CfsSubse-

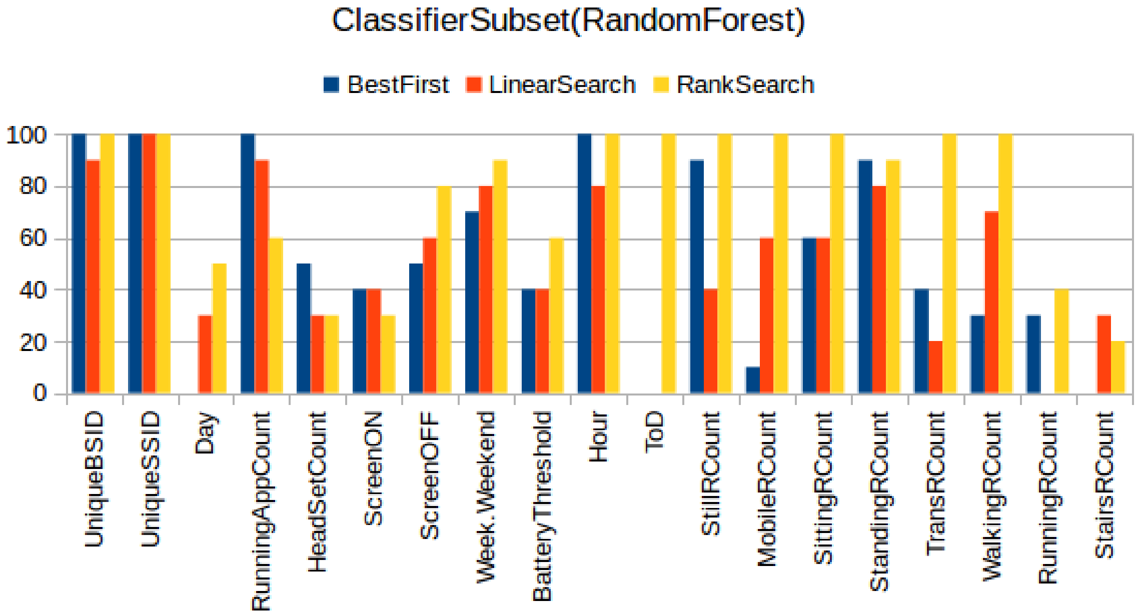


Figure 4.15: ClassifierSubsetEval with Random Forest Results for All Participants

tEval method in Figure 4.14, for Random Forest in Figure 4.15, for decision tree in Figure 4.16 and in Figure 4.17 for Naive Bayes algorithm. Additionally, in Figure 4.19, we present the average number of times each feature is selected by all 5 algorithms

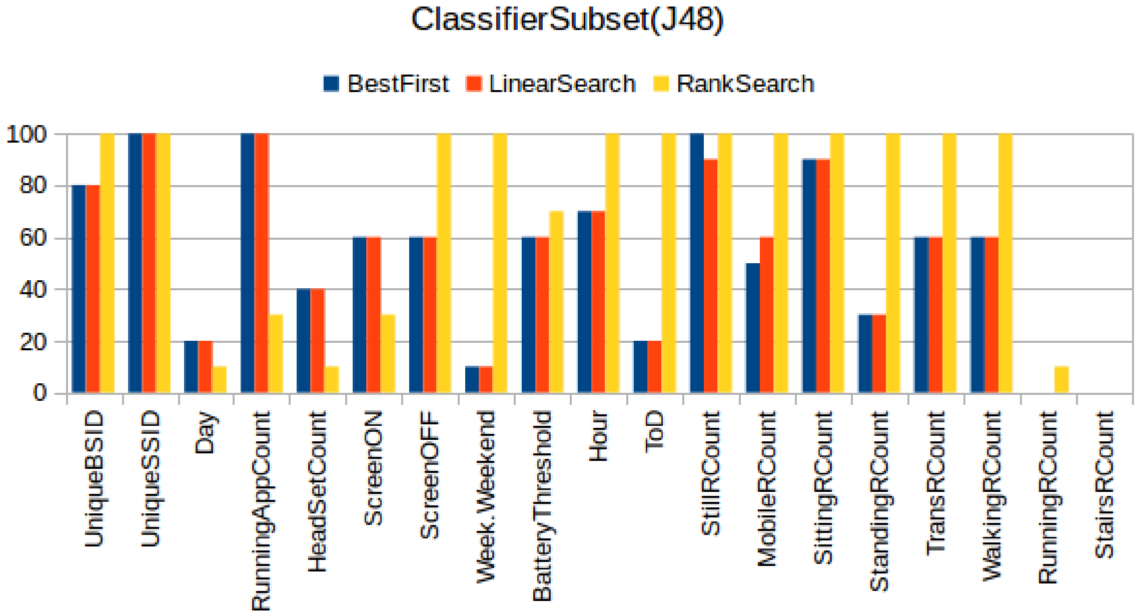


Figure 4.16: ClassifierSubsetEval with J48 Decision Tree Results for All Participants

and 3 search methods.

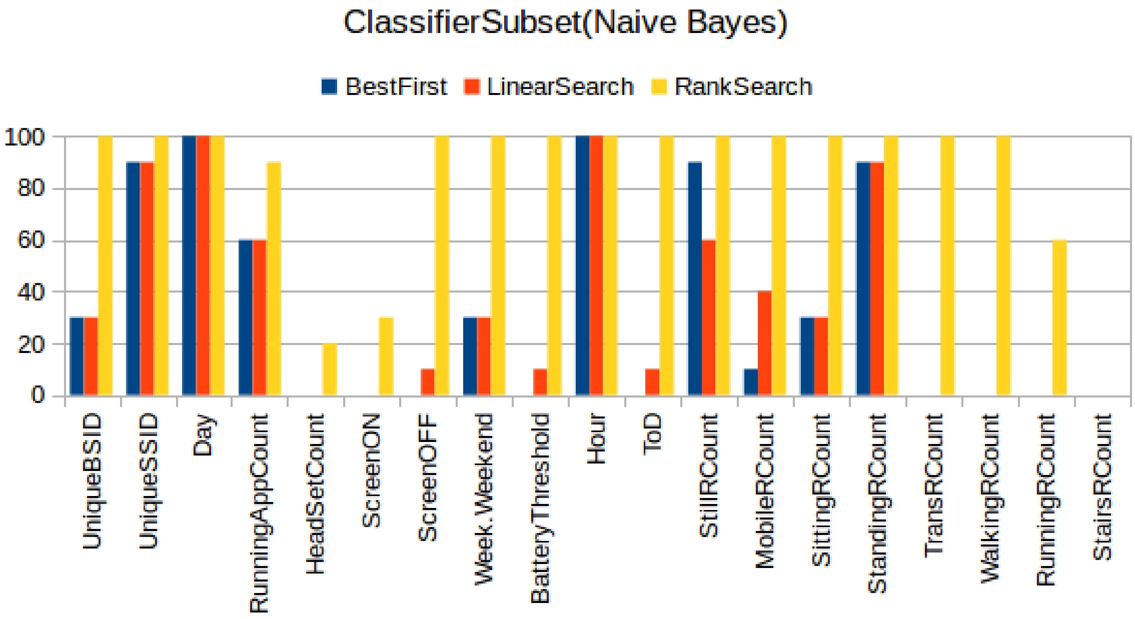


Figure 4.17: ClassifierSubsetEval with Naive Bayes Results for All Participants

While StairRCount and RunningRCount Features returned low rates for all the four options, UniqueBSID and UniqueSSID Feature returned high rates. StandingRCount Feature is a good option to choose according to all the four methods.

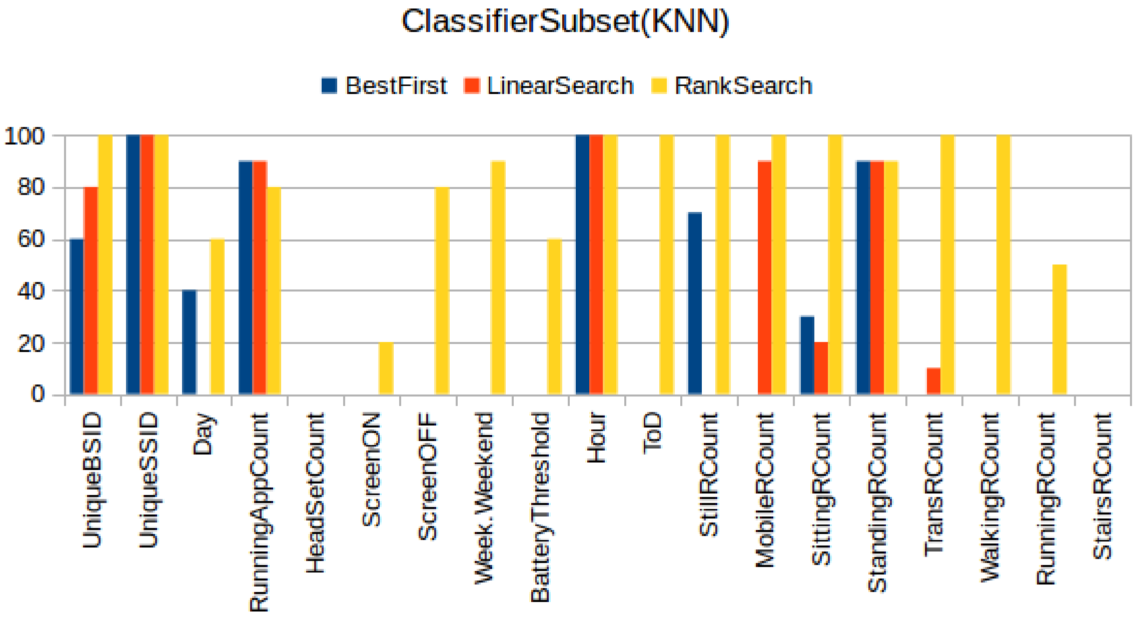


Figure 4.18: ClassifierSubsetEval with K-Nearest Neighbour Results for All Participants

Day Feature is one of the first attributes identified for data classification in the pre-

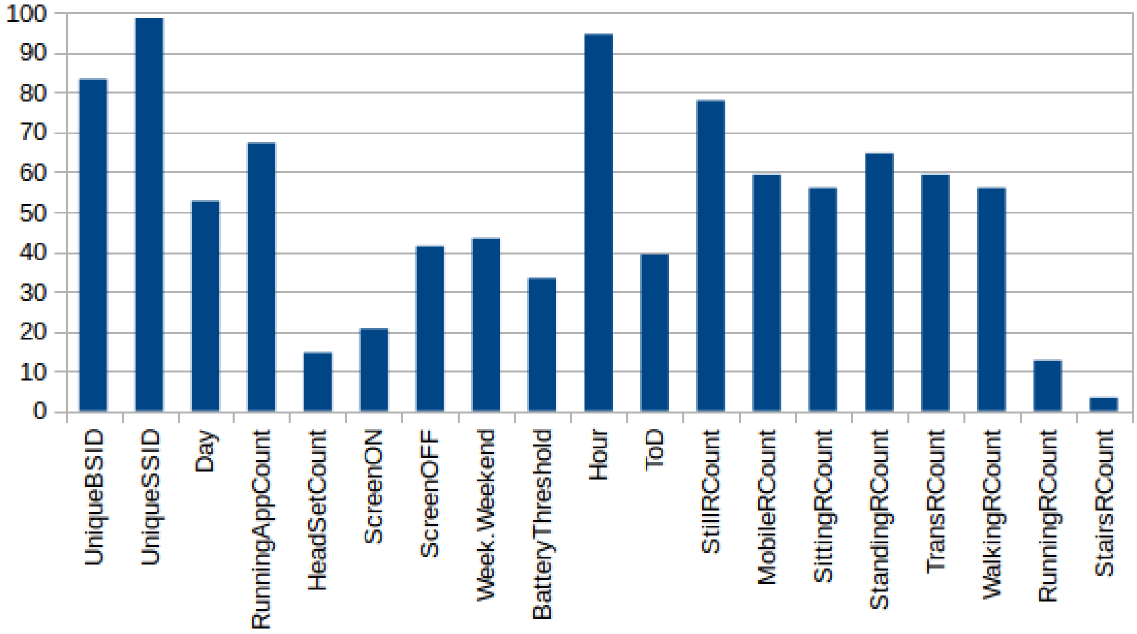


Figure 4.19: Average Nr. of Times (in %) Each Feature Selected by All 5 Algorithms and 3 Search Methods

liminary stages. Even though we have expected high results, it seems CfsSubsetEval selection algorithm and ClassifierSubsetEval algorithms with Naive Bayes favour this feature. As you can see in Figure 4.14 both BestFirst and LinearSearch methods picked it with 100% rate. For the ClassifierSubsetEval - Naive Bayes combination, Day Feature is chosen with 100% rate with all three search methods. Unlike the Day Feature, Hour Feature was selected by all the algorithms. And with the exception of ClassifierSubsetEval - Random Forest combination and ClassifierSubsetEval - J48 decision tree combination, the Hour Feature returns 100% rate. (Figures 4.16 and 4.15)

ToD Feature is highly picked only with RankSearch method among all the selection algorithms. ClassifierSubsetEval with Random Forest algorithm selects all features with varying degrees. Especially combined with RankSearch method even the normally not picked features returned least 20% rates. StairsRCount was not chosen with any other selection algorithms we used (Figure 4.15).

As a summary, in Figure 4.19, we observe that UniqueSSID, Hour, UniqueBSID, StillCount are the mostly selected features by the algorithms. They are selected more than 80% of the time. RunningAppCount, StandingCount, MobileCount, TransCount, SittingCount and WalkingCount were also selected more than 50% times. StairsCount and RunningCount have very low rates since ARService cannot predict these activities accurately. Week/Weekend feature was seldomly selected since the number of instances collected in our dataset at the weekends were very low. Similarly ScreenON, ScreenOFF and HeadsetCount features were not selected most of the time. We were expecting that screen features could give us hints about the participants' interaction time with the phone and this may change from place to place. However, this was not observed in our dataset. We plan to explore other phone-related features as a future work. Similarly, another phone feature, battery level, was not among the mostly selected features. We guess that this is due to the fact that participants did not have regular charging patterns, such as they charge it when they go home. As for the Activity Features, StillRCount is the best selection of our data set. StandingRCount, MobileRCount and TransRCount are the other good options.

4.5 Impact of Validation

In this section, our aim is to analyze the data in two different methods : person-dependent analysis where we use both the training data and the test data only from a specific person and person-independent analysis where we use the leave-one-subject-out method, such that when we use test data from a specific participant, his/her data is not available in the training set. Compared to our analysis in Section 4.3, where we combined all the data and used cross validation, in this section, our aim is to analyze the impact of personalization.

4.5.1 Person Dependent Results

In this section, we process the data in a person-dependent manner : We assume that both the training data and test data are from the same user. We analyze the performance with different feature combinations similar to Section 4.3.

Results with J48, decision tree, random forest, KNN and naive Bayes algorithms are given in Figure 4.20. All the data from stage-3 participants for one month duration is processed. After removing duplicate data, and removing mismatches in labelling the remaining data is classified with using Time, Communication, Activity and Phone

Features.

Person dependent classification results for our participants mostly have around 60% success rate. However the participants 1 and 2 returned the lowest rates with K-Nearest Neighbour classification algorithm. The success rate decreased to 30% for the participant 2. Random Forest classification algorithm is one of the best options according to Figure 4.20 with 50% rate. With the exception of the participant 7, which returned around 40% rate.

The participant 13 results are the odd one out within the results. Despite being the lowest label number (only three label tagged by participant 13), the prediction success rate is around 90%.

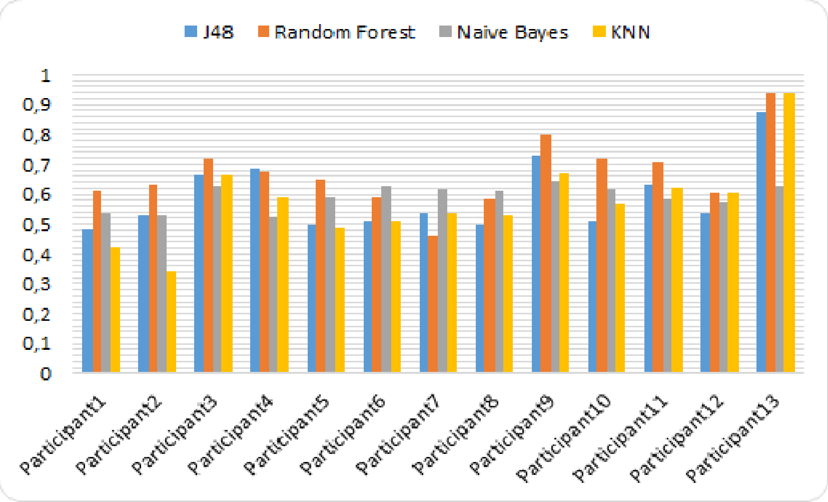


Figure 4.20: Person Dependent Classification using All Features with All Classification Algorithms

4.5.2 Person-Independent Results

In this section, we present the results when the data is processed in a person-independent manner, such that when a specific participant is selected for testing/classification, his/-her data is excluded from the training set. The effect of Leave One Subject Out can be observed by comparing the Cross Validation results. One of the major differences between 4.12 and 4.21 is the number of elements in x-axis. Since two of the participants (Participant 7 and Participant 13) had marked only small number of labels, they returned relatively high scores (almost 100% correct prediction), they are not included at the LOSO classifications as a separate test data.

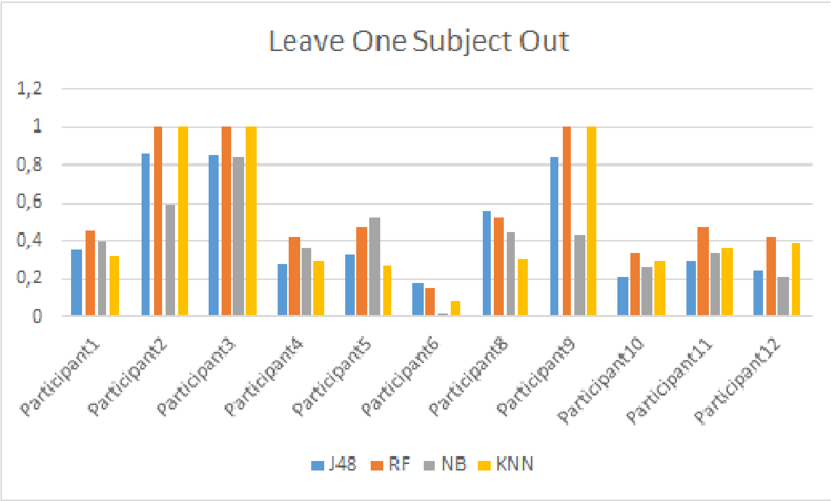


Figure 4.21: Leave One Subject Out Classification Results

When we observe the results, Random Forest and K-Nearest Neighbour classification algorithms delivered the highest prediction rates (100%) with Participants 2,3 and 9. Random Forest tends to return high rates for all participants while K-Nearest Neighbour algorithm changes from test set to test set. For participant 1 and 8 the K-Nearest Neighbour algorithm returned the lowest results of the four classification algorithms.

When we compare these results with person dependent results given in Figure 4.20,

while participants 3 and 9 return high results both in person dependent and person independent classification, however we cannot say the same for other results. One of the most drastic changes is the change of Participant 6' s results. With person independent classification, we obtained success rates between 50 to 60%. However when we processed with the LOSO method, the success rates decrease to 20%.

On the other hand for the participants 2,3 and 9 person-dependent results return 100% success with Random Forest and K-Nearest Neighbour algorithms. All the other participants' results drop down at least by 10%. In the person independent classification, Naive Bayes and KNN algorithms return the lowest success rates for almost all participants. For participants 1, 2, 5, 6 and 10 KNN returns the lowest rates, while the participants 3,9,11 and 13 got lowest rates with Naive Bayes. Random Forest algorithm has the highest success rate among other algorithms. The detailed information per each class is presented at the Appendix.

One of the other reasons in terms of difference between person dependent and independent results is due to the variety of the labels tagged by the participants. Some participants used all the predefined labels in their daily routines, while others only used 3 or 4 labels to mark their places. Participant 13 only used "Home", "Work/School" and "Canteen/Restaurant/Cafe/Bar" as the place labels. There were differences in the results between person dependent and person independent classification for this participant. Since there were only 3 possibilities to choose from, the person dependent search (cross validation) returns high prediction rates with 100 % rate. For this particular case, person independent search (leave-one-subject-out) the prediction rates were lower. Since the total number of place labels are 11, the participants with lower tag numbers return lower results. However, for the rest of the participant data ; the difference between tagged label number and total label number did not make a particular difference. The total number of tagged labels also have a non- uniform distribution. "On the Road (Transportation)" label were tagged 784 times, while "[GSU] Library" only tagged 8 times as shown in Figure 4.1

5 CONCLUSION

The studies on semantic place tagging in the literature preferred to use GPS, or social media tracking information. Even though it returns high correct prediction rates, it creates privacy, data package and battery problems.

In this research we preferred to use a more anonymous approach. Using total number of connected devices, instead of devices' ids, leaving GPS' langtitude, longtitude data altogether, sending collected data only when the smart phone is connected to a WiFi are our solutions.

In this thesis, we collected mobile phone data from 20 participants from Galatasaray University for a month-duration using a sensor logger application named ARService. These data include information about the communication information, such as BSID' s and SSID' s from which the phone receives signals, phone-usage information, such as the number of running applications, battery level, activity features, whether the person is still/mobile, walking, in transportation mode, etc. Besides these information automatically collected from the phone, we also used information about time. Since people have daily routines in life, we used hour, time of the day (morning, afternoon, etc.) and whether a day is a weekday or a weekend since people's visiting patterns change at the weekends.

Before processing the data, first we visualized how different features correlate with different places. Next, we explored the impact of using different feature sets, related to time, communication, phone usage and activities, on identifying such places. We observe that, when all features are used together better recognition accuracies are achieved. However, the combination of activity, time and communication information also provide very close results compared with using all features.

Afterwards, we analyzed the impact of each feature instead of feature sets. Similarly, we observe that UniqueSSID (communication), Hour (time), UniqueBSID (communication), StillCount (activity) are the mostly selected features by five different feature selection algorithms combined with three different search methods. Attribute selection methods reveals Hour feature has more impact than ToD feature. Even though

ToD (Time of the Day) feature created grouping Hour Feature, it becomes less effective. Walking Feature has on of the least important attributes in the Activity Features group. While Sitting has considerably more effective on our results.

In the last part of our analysis, we focused on whether person-independent classification can achieve high rates compared to a person-dependent approach and showed that if the participant has a lot of training data, than a person-dependent approach can achieve higher rates. However, if the participant does not have labelled data, then using training data from other people can achieve good rates. In all steps of our analysis, we also explored the performance with four different classification algoritms and random forest classifier achieved the best results.

As for future work of this study we are planning to use SVM classification algorithm with collected data of stage three. Analysing behaviour of people through seasons with our base four participants data is also another thing we plan to do. Detailing the difference caused by age and gender is part of our plans.

REFERENCES

Berchtold, M., Budde, M., Gordon, D., Schmidtke, H. and Beigl, M. (2010). Actiserv : Activity recognition service for mobile phones, *Wearable Computers (ISWC), 2010 International Symposium on*, pp. 1 –8.

Çelik, S. C. and İncel, Ö. D. (2016). Semantic place prediction from mobile phone sensors, *2016 24th Signal Processing and Communication Application Conference (SIU)*, IEEE, pp. 1021–1024.

Coskun, D. (2014). *Real-time activity recognition on smart phones*, Master's thesis, Galatasaray University.

Do, T. M. T. and Gatica-Perez, D. (2014). The places of our lives : Visiting patterns and automatic labeling from longitudinal smartphone data, *Mobile Computing, IEEE Transactions on* **13**(3) : 638–648.

Eagle, N. and Pentland, A. (2006). Reality mining : sensing complex social systems, *Personal and ubiquitous computing* **10**(4) : 255–268.

Gütlein, M., Frank, E., Hall, M. and Karwath, A. (2009). Large-scale attribute selection using wrappers, *Computational Intelligence and Data Mining, 2009. CIDM'09. IEEE Symposium on*, IEEE, pp. 332–339.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*, PhD thesis, The University of Waikato.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009). The weka data mining software : an update, *ACM SIGKDD explorations newsletter* **11**(1) : 10–18.

Kira, K. and Rendell, L. A. (1992). A practical approach to feature selection, *Proceedings of the ninth international workshop on Machine learning*, pp. 249–256.

Krumm, J. and Rouhana, D. (2013). Placer : semantic place labels from diary data, *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, ACM, pp. 163–172.

Krumm, J., Rouhana, D. and Chang, M.-W. (2015). Placer++ : Semantic place labels beyond the visit, *Pervasive Computing and Communications (PerCom), 2015 IEEE International Conference on*, IEEE, pp. 11–19.

Laurila, J. K., Gatica-Perez, D., Aad, I., Blom, J., Bornet, O., Do, T. M. T., Dousse, O., Eberle, J. and Miettinen, M. (2013). From big smartphone data to worldwide research : the mobile data challenge, *Pervasive and Mobile Computing* **9**(6) : 752–771.

Laurila, J. K., Gatica-Perez, D., Aad, I., Bornet, O., Do, T.-M.-T., Dousse, O., Eberle, J., Miettinen, M. et al. (2012). The mobile data challenge : Big data for mobile computing research, *Pervasive Computing*. Paper no. EPFL-CONF-192489.

Li, G., Yu, L., Ng, W. S., Wu, W. and Goh, S. T. (2015). Predicting home and work locations using public transport smart card data by spectral analysis, *Intelligent Transportation Systems (ITSC), 2015 IEEE 18th International Conference on*, IEEE, pp. 2788–2793.

Lian, D., Zhu, Y., Xie, X. and Chen, E. (2014). Analyzing location predictability on location-based social networks, *Advances in Knowledge Discovery and Data Mining*, Springer, pp. 102–113.

Madan, A., Cebrian, M., Moturu, S., Farrahi, K. and Pentland, A. (2012). Sensing the" health state" of a community, *IEEE Pervasive Computing* **11**(4) : 36–45.

RStudio Team (2015). *RStudio : Integrated Development Environment for R*, RStudio, Inc., Boston, MA.
URL: *http ://www.rstudio.com/*

Zhu, Y., Zhong, E., Lu, Z. and Yang, Q. (2013). Feature engineering for semantic place prediction, *Pervasive and mobile computing* **9**(6) : 772–783.

A APRIL PARTICIPANT RESULTS

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	0.762	0.286	0.619	0.667	0.571	0.667	0.667	0.619	0.333	0.571	0.429
Outdoor (Park, etc.)	0	0	0	0.444	0.111	0.444	0.444	0.444	0	0.444	0
On the Road (Transportation)	0.516	0.613	0.452	0.548	0.419	0.548	0.645	0.677	0.839	0.355	0.452
[GSU] Canteen/garden	0.444	0.111	0.389	0.5	0.389	0.5	0.222	0.222	0	0.222	0.278
Mall/Shop	0	0	0.188	0.313	0.063	0.375	0.375	0.375	0	0.375	0.063
Canteen/Restaurant/Cafe/Bar	0	0	0	0	0	0	0	0	0	0	0
[GSU] Classroom/Lab	0	0.111	0.111	0.556	0.444	0.556	0.222	0.333	0.222	0	0
Weighted Average	0.377	0.264	0.358	0.509	0.358	0.519	0.472	0.481	0.33	0.3349	0.274

Table A.1: Participant1 J48 Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	0.429	0.524	0.429	0.762	0.524	0.714	0.81	0.81	0.333	0.524	0.333
Outdoor (Park, etc.)	0	0	0	0.111	0	0.22	0.222	0.222	0	0.444	0.222
On the Road (Transportation)	0.226	0.387	0.355	0.613	0.419	0.677	0.71	0.742	0.581	0.677	0.581
[GSU] Canteen/garden	0.444	0.333	0.333	0.333	0.333	0.389	0.389	0.44	0.278	0.278	0.278
Mall/Shop	0	0.063	0.125	0.375	0.125	0.375	0.375	0.5	0.063	0.438	0.25
Canteen/Restaurant/Cafe/Bar	0.5	1	1	1	1	1	1	1	1	1	1
[GSU] Classroom/Lab	0.222	0.333	0.111	0.556	0.333	0.556	0.556	0.556	0.222	0	0.444
Weighted Average	0.255	0.33	0.292	0.519	0.349	0.547	0.575	0.613	0.33	0.472	396

Table A.2: Participant1 Random Forest Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	0.524	0.381	0.667	0.762	0.619	0.619	0.81	0.857	0.333	0.476	0.667
Outdoor (Park, etc.)	0	0	0	0.222	0	0.111	0.111	0.111	0	0.222	0
On the Road (Transportation)	0.484	0.581	0.258	0.387	0.323	0.548	0.387	0.419	0.742	0.323	0.226
[GSU] Canteen/garden	0.556	0.278	0.333	0.5	0.333	0.556	0.5	0.5	0	0.389	0.167
Mall/Shop	0.125	0	0.438	0.563	0.375	0.438	0.563	0.438	0	0.688	0.313
Canteen/Restaurant/Cafe/Bar	0	0	1	1	1	1	1	1	1	1	1
[GSU] Classroom/Lab	0.111	0.444	0.444	0.556	0.778	0.778	0.778	0.778	0.889	0.333	0.444
Weighted Average	0.368	0.33	0.387	0.519	0.415	0.538	0.538	0.538	0.377	0.425	0.33

Table A.3: Participant1 KNN Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	0.429	0.524	0.476	0.667	0.476	0.619	0.571	0.571	0.381	0.524	0.476
Outdoor (Park, etc.)	0	0	0	0	0	0	0	0.	0	0.444	0.333
On the Road (Transportation)	0.323	0.355	0.29	0.419	0.323	0.452	0.452	0.419	0.548	0.613	0.613
[GSU] Canteen/garden	0.389	0.278	0.333	0.444	0.278	0.444	0.5	0.5	0.167	0.333	0.389
Mall/Shop	0	0.125	0.125	0.313	0.125	0.25	0.313	0.313	0	0.375	0.438
Canteen/Restaurant/Cafe/Bar	0	1	1	1	1	1	1	1	1	1	1
[GSU] Classroom/Lab	0.222	0.333	0.222	0.444	0.222	0.444	0.444	0.444	0.222	0.111	0.444
Weighted Average	0.264	0.321	0.292	0.434	0.292	0.425	0.434	0.425	0.302	0.462	0.491

Table A.4: Participant1 NB Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Work/School	0	0.5	0.5	0	0.5	0	0.5	0	0	0	1
Home	0.451	0.647	0.569	0.314	0.686	0.588	0.431	0.549	0.558	0.392	0.471
On the Road (Transportation)	0.28	0.42	0.3	0.3	0.480	0.560	0.28	0.48	0.38	0.392	0.22
[GSU] Canteen/garden	0.286	0.531	0.469	0.531	0.469	0.490	0.469	0.551	0.469	0.469	0.429
Canteen/Restaurant/Cafe/Bar	0	0.333	0	0.143	0.333	0.333	0.048	0.238	0	0.143	0
[GSU] Classroom/Lab	0.263	0.711	0.289	0.474	0.526	0.658	0.421	0.526	0.684	0.316	0.289
[GSU] Library	0	0	0	0	0	0	0	0	0	0	0
Outdoor (Park, etc.)	0.116	0.302	0.233	0.744	0.349	0.744	0.767	0.744	0.209	0.698	0.256
Mall/Shop	0	0	0	0	0.486	0.568	0.428	0.529	0	0	0
Weighted Average	0.257	0.498	0.346	0.428	0.292	0.425	0.434	0.425	0.416	0.409	0.304

Table A.5: Participant2 J48 Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Work/School	0	0.5	1	1	1	0.5	1	1	0	1	1
Home	0.353	0.627	0.608	0.392	0.804	0.745	0.529	0.765	0.529	0.314	0.529
On the Road (Transportation)	0.38	0.46	0.44	0.28	0.540	0.6	0.340	0.54	0.44	0.28	0.38
[GSU] Canteen/garden	0.306	0.388	0.449	0.49	0.531	0.551	0.510	0.612	0.408	0.531	0.408
Canteen/Restaurant/Cafe/Bar	0.048	0.333	0.095	0	0.143	0.238	0.143	0.19	0	0.286	0.238
[GSU] Classroom/Lab	0.395	0.579	0.342	0.526	0.553	0.737	0.447	0.763	0.684	0.237	0.368
[GSU] Library	0	0	0	0	0	0	0	0	0	0	0
Outdoor (Park, etc.)	0.116	0.302	0.209	0.674	0.302	0.721	0.721	0.721	0.488	0.744	0.233
Mall/Shop	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.284	0.455	0.393	0.424	0.518	0.623	0.475	0.63	0.451	0.409	0.377

Table A.6: Participant2 RF Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Work/School	0	0	0.5	1	1	1	1	1	0	1	1
Home	0.51	0.529	0.431	0.137	0.549	0.373	0.216	0.353	0.431	0.02	0.49
On the Road (Transportation)	0.08	0.4	0.18	0.12	0.360	0.42	0.16	0.46	0.38	0.12	0.18
[GSU] Canteen/garden	0.592	0.612	0.673	0.49	0.612	0.551	0.429	0.531	0.551	0.388	0.673
Canteen/Restaurant/Cafe/Bar	0	0.333	0	0.524	0.238	0.476	0.476	0.476	0	0.19	0
[GSU] Classroom/Lab	0.105	0.684	0.053	0.737	0.474	0.737	0.632	0.632	0.816	0.533	0
[GSU] Library	0	0	0	0	0	0	0	0	0	0	0
Outdoor (Park, etc.)	0.023	0.186	0.163	0.814	0.256	0.791	0.814	0.791	0.209	0.814	0.163
Mall/Shop	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.249	0.459	0.288	0.44	0.436	0.549	0.432	0.533	0.42	0.342	0.296

Table A.7: Participant2 NB Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Work/School	0	0.5	1	1	1	1	1	1	0	1	1
Home	0.49	0.471	0.51	0.216	0.588	0.235	0.392	0.451	0.569	0.275	0.431
On the Road (Transportation)	0.46	0.4	0.34	0.26	0.360	0.3	0.26	0.38	0.44	0.169	0.32
[GSU] Canteen/garden	0.286	0.429	0.367	0.347	0.408	0.286	0.224	0.224	0.449	0.388	0.265
Canteen/Restaurant/Cafe/Bar	0	0.19	0.143	0.048	0.19	0.095	0.095	0.095	0	0.381	0.238
[GSU] Classroom/Lab	0.263	0.447	0.237	0.395	0.289	0.526	0.289	0.342	0.684	0.184	0.263
[GSU] Library	0	0	0	0	0	0	0	0	0	0	0
Outdoor (Park, etc.)	0.07	0.326	0.209	0.465	0.163	0.488	0.512	0.535	0.419	0.721	0.279
Mall/Shop	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.292	0.393	0.327	0.307	0.358	0.355	0.315	0.342	0.455	0.374	0.311

Table A.8: Participant2 KNN Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
On the Road (Transportation)	0.884	0.625	0.75	0.75	0.719	0.906	0.906	0.969	0.813	0.719	0.75
Outdoor (Park, etc.)	0	0	0	0.75	0	0.75	0.75	0.75	0	0.25	0
Canteen/Restaurant/Cafe/Bar	0.444	0.444	0.556	0.556	0.444	0.333	0.333	0.333	0.556	0.333	0
Work/School	0	0	0	0.167	0.333	0	0	0.167	0	0.333	0
Home	0.5	0.55	0.4	0.5	0.45	0.75	0.6	0.6	0.75	0.5	0.45
[GSU] Classroom/Lab	0	0	0	0	0	0	0	0	0	0	0
[GSU] Canteen/garden	0	0	0	0	0	0	0	0	0	0	0
Other	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.547	0.467	0.493	0.573	0.507	0.667	0.627	0.667	0.613	0.52	0.44

Table A.9: Participant3 J48 Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
On the Road (Transportation)	0.75	0.781	0.875	0.656	0.844	1	0.969	0.969	0.781	0.813	0.719
Outdoor (Park, etc.)	0.25	0	0.5	0.5	0.25	0.75	0.75	0.75	0	0	0.5
Canteen/Restaurant/Cafe/Bar	0.556	0.444	0.667	0	0.556	0.444	0.556	0.556	0.556	0.111	0.333
Work/School	0	0.833	0	0.833	0.167	0.167	0	0.167	0.167	0	0
Home	0.5	0.55	0.6	0.55	0.55	0.7	0.7	0.7	0.6	0.45	0.55
[GSU] Classroom/Lab	0	0	0	0	0	0	0	0	0	0	0
[GSU] Canteen/garden	0	0	0	0	0	0	0	0	0	0	0
Other	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.533	0.56	0.64	0.52	0.6	0.72	0.693	0.72	0.573	0.177	0.24

Table A.10: Participant3 RF Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
On the Road (Transportation)	0.906	0.781	0.813	0.656	0.781	0.656	0.688	0.813	0.656	0.531	0.75
Outdoor (Park, etc.)	0.75	0.25	0.5	0.5	0.25	0.25	0.5	0.25	0	0.25	0.5
Canteen/Restaurant/Cafe/Bar	0.333	0.667	0.444	0	0.444	0.111	0.111	0.111	0	0	0.111
Work/School	0	0	0	0.833	0	0.667	0.833	0.667	0	0.833	0
Home	0.35	0.7	0.6	0.55	0.7	0.55	0.7	0.75	0.8	0.35	0.35
[GSU] Classroom/Lab	0	0	0	0	0	0	0	0	0	0	0
[GSU] Canteen/garden	0	0	0	0	0	0	0	0	0	0	0
Other	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.56	0.613	0.587	0.52	0.587	0.507	0.587	0.627	0.493	0.4	0.453

Table A.11: Participant3 NB Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
On the Road (Transportation)	0.813	0.688	0.75	0.813	0.75	0.781	0.813	0.813	0.688	0.813	0.719
Outdoor (Park, etc.)	0.5	0	0.5	0.75	0.5	0.75	0.75	0.75	0	0.75	0.5
Canteen/Restaurant/Cafe/Bar	0.556	0.444	0.778	0.667	0.667	0.667	0.667	0.778	0.556	0.222	0.444
Work/School	0	0.333	0	0	0	0.333	0	0	0	0.333	0
Home	0.35	0.45	0.5	0.7	0.55	0.65	0.7	0.7	0.65	0.45	0.35
[GSU] Classroom/Lab	0	0	0	0	0	0	0	0	0	0	0
[GSU] Canteen/garden	0	0	0	0	0	0	0	0	0	0	0
Other	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.533	0.493	0.573	0.653	0.573	0.653	0.653	0.667	0.533	0.56	0.48

Table A.12: Participant3 KNN Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	0.762	0.857	0.714	0.81	0.81	0.857	0.81	0.81	0.81	0.571	0.429
On the Road,(Transportation)	0.55	0.6	0.65	0.5	0.55	0.6	0.55	0.7	0.4	0.4	0.3
Work/School	1	0.333	0.333	0.667	0	0.333	0.333	0.333	0	0	0
[GSU]Classroom/Lab	0.925	0.925	0.85	0.925	0.825	0.925	0.85	0.85	0.85	0.625	0.725
Other	0	0	0	0	0	0	0	0	0	0	0
[GSU] Canteen/garden	0	0.25	0	0.125	0	0.25	0.375	0.25	0	0.125	0
Gym	0	0	0	0	0	0	0	0	0	0	0
Outdoor (Park, etc.)	0	0	0	0	0	0	0	0	0	0	0
Mall/Shop	1	0.75	0.5	0.75	1	0.75	0.75	1	1	0.75	0.25
Canteen/Restaurant/Cafe/Bar	0	0	0	0	0	0	0	0	0	0	0
[GSU] Library	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.676	0.695	0.619	0.667	0.619	0.695	0.657	0.686	0.6	0.467	0.429

Table A.13: Participant4 J48 Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	0.857	0.905	0.857	0.81	0.905	0.905	0.381	0.857	0.81	0.524	0.571
On the Road,(Transportation)	0.75	0.6	0.6	0.6	0.5	0.6	0.4	0.6	0.4	0.5	0.35
Work/School	0.667	0.333	0.333	0	1	0.333	0	0.333	0	0	0.333
[GSU]Classroom/Lab	0.85	0.85	0.8	0.825	0.925	0.85	0.675	0.9	0.7	0.65	0.625
Other	0	0	0	0	0	0	0	0	0	0	0
[GSU] Canteen/garden	0	0	0	0.125	0	0	0	0	0	0	0
Gym	0	0	0	0	0	0	0	0	0	0	0
Outdoor (Park, etc.)	0	0	0	0	0	0	0	0	0	0	0
Mall/Shop	0.5	1	0.75	0.75	1	1	0.75	1	1	0.75	0.5
Canteen/Restaurant/Cafe/Bar	0	0	0	0	0	0	0	0	0	0	0
[GSU] Library	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.676	0.081	0.667	0.638	0.695	0.667	0.476	0.676	0.543	0.467	0.467

Table A.14: Participant4 RF Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	0.524	0.571	0.381	0.381	0.667	0.571	0.381	0.381	0.095	0	0.048
On the Road,(Transportation)	0.05	0.6	0.4	0.3	0.45	0.6	0.65	0.65	0.3	0.3	0.2
Work/School	0	0	0.333	0	0	0	0	0	0	0	0.333
[GSU]Classroom/Lab	0.85	0.825	0.35	0.8	0.325	0.825	0.7	0.7	0.625	0.525	0.175
Other	0	0	0	0	0	0	0	0	0	0	0
[GSU] Canteen/garden	0	0.25	0.375	0	0.5	0.25	0.375	0.375	0	0	0.5
Gym	0	0	0	0	0	0	0	0	0	0	0
Outdoor (Park, etc.)	0	0	0	0	0	0	0	0	0	0	0
Mall/Shop	0.75	0.75	0.5	0.75	0.75	0.75	0.75	0.75	1	0.25	0.25
Canteen/Restaurant/Cafe/Bar	0	0	0	0	0	0	0	0	0	0	0
[GSU] Library	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.467	0.59	0.343	0.467	0.41	0.59	0.467	0.524	0.352	0.267	0.171

Table A.15: Participant4 NB Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	0.875	0.875	0.875	0.875	0.875	0.875	0.875	0.875	0.619	0.571	0.524
On the Road,(Transportation)	0.75	0.55	0.4	0.6	0.3	0.55	0.5	0.55	0.4	0.45	0.25
Work/School	1	0	1	0	0.667	0	0	0	0	0	0
[GSU]Classroom/Lab	0.85	0.825	0.75	0.7	0.825	0.825	0.7	0.75	0.65	0.575	0.625
Other	0	0	0	0	0	0	0	0	0	0	0
[GSU] Canteen/garden	0	0	0	0	0	0	0	0	0	0	0.25
Gym	0	0	0	0	0	0	0	0	0	0	0
Outdoor (Park, etc.)	0	0	0	0	0	0	0	0	0	0	0
Mall/Shop	0.25	0.75	0.5	0.75	0.75	0.75	0.75	0.75	1	0.75	0.5
Canteen/Restaurant/Cafe/Bar	0	0	0	0	0	0	0	0	0	0	0
[GSU] Library	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.676	0.619	0.581	0.581	0.59	0.619	0.562	0.59	0.486	0.448	0.429

Table A.16: Participant4 KNN Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
n theRoad (Transportation)	0.667	0.872	0.436	0.564	0.795	0.692	0.59	0.692	0.795	0.59	0.487
Work/School	0.963	0.438	0.313	0.375	0.438	0.313	0.425	0.313	0.5	0.25	0.375
Home	0.615	0.769	0.462	0.308	0.692	0.615	0.385	0.692	0.692	0.231	0.154
Canteen/Restaurant/Cafe/Bar	0.071	0.643	0.429	0.357	0.5	0.5	0.429	0.429	0.5	0.429	0.071
Mall/Shop	0	0	0	0	0	0	0	0	0	0.25	0
Other	0	0	0.2	0.4	0	0	0	0	0.4	0	0.25
Gym	0	0	0	1	0	0	0	0.6	0	0.5	0
Outdoor(Park, etc.)	0.125	0	0.375	0.375	0	0.125	0.375	0.125	0.25	0.25	0
[GSU]Classroom/Lab	0	0	0	0	0	0	0	0	0	0	0
WeightedAverage	0.363	0.588	0.373	0.431	0.529	0.5	0.402	0.5			

Table A.17: Participant5 J48 Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
On theRoad (Transportation)	0.333	0.692	0.513	0.692	0.744	0.795	0.769	0.795	0.744	0.718	0.513
Work/School	0.5	0.563	0.438	0.375	0.5	0.5	0.438	0.5	0.375	0.125	0.375
Home	0.538	0.615	0.308	0.385	0.692	0.615	0.385	0.615	0.692	0.231	0
Canteen/Restaurant/Cafe/Bar	0.214	0.5	0.571	0.571	0.571	0.571	0.714	0.714	0.643	0.5	0.5
Mall/Shop	0	0	0	0.25	0	0.25	0.25	0.5	0.5	0.5	0.5
Other	0.2	0.4	0	0.4	0.4	0.4	0.2	0.4	0.4	0	0.2
Gym	0.5	1	1	1	1	1	1	1	1	0.5	0.5
Outdoor(Park, etc.)	0	0.25	0.375	0.25	0.375	0.375	0.25	0.375	0.5	0.375	0.5
[GSU]Classroom/Lab	0	0	0	0	0	0	0	0	0	0	0
WeightedAverage	0.324	0.559	0.431	0.52	0.598	0.618	0.569	0.175	0.618	0.451	0.392

Table A.18: Participant5 RF Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
On theRoad (Transportation)	0.667	0.718	0.564	0.564	0.744	0.744	0.513	0.821	0.769	0.487	0.41
Work/School	0.188	0.563	0.5	0.375	0.5	0.375	0.5	0.5	0.063	0.125	0.25
Home	0.538	0.692	0.769	0.615	0.769	0.769	0.692	0.769	0.846	0.615	0.154
Canteen/Restaurant/Cafe/Bar	0.214	0.286	0.214	0	0.286	0.214	0	0.214	0.643	0	0.143
Mall/Shop	0	0	0	0	0	0	0.5	0.25	0	0	0
Other	0	0.4	0	0.2	0.2	0.6	0	0.2	0.2	0.2	0
Gym	0.5	1	0.5	0	0.5	0	0	0	1	0	0
Outdoor(Park, etc.)	0	0.125	0.125	0.75	0.125	0.75	0.625	0.625	0.25	0.75	0
[GSU]Classroom/Lab	0	0	0	0	0	0	0	0	0	0	0
WeightedAverage	0.392	0.539	0.441	0.422	0.529	0.559	0.431	0.588	0.549	0.353	0.235

Table A.19: Participant5 NB Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
On theRoad (Transportation)	0.487	0.641	0.333	0.538	0.615	0.615	0.513	0.59	0.564	0.538	0.462
Work/School	0.438	0.438	0.5	0.438	0.438	0.375	0.438	0.385	0.313	0.125	0.25
Home	0.385	0.692	0.308	0.231	0.462	0.308	0.308	0.643	0.538	0.154	0.154
Canteen/Restaurant/Cafe/Bar	0.214	0.429	0.429	0.571	0.429	0.571	0.643	0	0.643	0.571	0.5
Mall/Shop	0	0	0	0	0	0	0	0.25	0.5	0.5	0.5
Other	0	0.4	0	0.6	0	0.6	0	0.2	0.4	0	0
Gym	0.5	1	1	1	1	1	1	1	0.5	1	0
Outdoor(Park, etc.)	0	0.625	0.25	0.5	0.25	0.5	0.5	0.5	0.375	0.375	0.375
[GSU]Classroom/Lab	0	0	0	0	0	0	0	0	0	0	0
WeightedAverage	0.343	0.549	0.343	0.471	0.461	0.5	0.451	0.49	0.5	0.392	0.353

Table A.20: Participant5 KNN Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
[GSU]Canteen/garden	0.083	0.25	0.333	0.25	0.333	0.333	0.5	0.5	0.583	0.667	0.25
[GSU]Classroom/Lab	0.708	0.667	0.75	0.625	0.667	0.625	0.667	0.667	0.75	0.75	0.55
Home	0	0	0.25	0.25	0.5	0.25	0.5	0.5	0	0.75	0.25
Work/School	0	0.333	0	0.333	0.167	0.333	0.167	0.333	0	0.333	0
Outdoor (Park, etc.)	0	0	0	0.333	0.333	0	0.333	0	0	0.333	0
Other	0	0	0	0	0	0	0	0	0	0	0
WeightedAverage	0.353	0.412	0.451	0.431	0.471	0.431	0.51	0.51	0.49	0.627	0.389

Table A.21: Participant6 J48 Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
[GSU]Canteen/garden	0.25	0.5	0.417	0.417	0.583	0.5	0.5	0.5	0.5	0.667	0.333
[GSU]Classroom/Lab	0.583	0.667	0.75	0.708	0.75	0.833	0.75	0.792	0.583	0.667	0.667
Home	0	0.25	0.25	0.5	0.5	0.25	0.5	0.5	0.5	0.5	0.75
Work/School	0	0.333	0.333	0.667	0.333	0.5	0.333	0.333	0.333	0.333	0.167
Outdoor (Park, etc.)	0	0	0	0.333	0	0.333	0.333	0.333	0.333	0.333	0
Other	0	0	0.5	0	0	0	0	0	0	0	0
WeightedAverage	0.353	0.49	0.529	0.596	0.569	0.608	0.569	0.588	0.49	0.569	0.471

Table A.22: Participant6 RF Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
[GSU]Canteen/garden	0.167	0.333	0.333	0.5	0.417	0.583	0.583	0.667	0.333	0.583	0.333
[GSU]Classroom/Lab	0.792	0.833	0.75	0.5	0.75	0.458	0.667	0.792	0.75	0.208	0.792
Home	0.5	0	0.75	0.75	0.5	0.5	0.75	0.5	0.25	0.5	0.5
Work/School	0	0.167	0.333	0.333	0.333	0.333	0.5	0.5	0.167	0.333	0.167
Outdoor (Park, etc.)	0	0	0	0	0	0	0	0	0	0	0
Other	0	0	0	0	0	0	0	0	0	0	0
WeightedAverage	0.451	0.49	0.529	0451	0.529	0.5431	0.569	0.627	0.471	0.314	0.293

Table A.23: Participant6 NB Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
[GSU]Canteen/garden	0.417	0.417	0.333	0.167	0.333	0.167	0.333	0.333	0.5	0.5	0.5
[GSU]Classroom/Lab	0.583	0.583	0.667	0.542	0.625	0.625	0.583	0.625	0.542	0.5	0.583
Home	0.5	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.5	0.5	0.75
Work/School	0	0.333	0.333	0.667	0.333	0.5	0.333	0.333	0.333	0.333	0.167
Outdoor (Park, etc.)	0	0.333	0	0.333	0	0	0.333	0.667	0.333	0	0
Other	0	0	0	0	0	0	0	0	0	0	0
WeightedAverage	0.412	0.49	0.49	0451	0.471	0.49	0.471	0.51	0.471	0.431	0.24

Table A.24: Participant6 KNN Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Work/School	0.615	0.667	0.833	0.833	0.667	0.667	0.833	0.667	0.667	0.333	0.5
Home	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	1	0.167	0.5
[GSU]Library	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.615	0.538	0.615	0.615	0.538	0.538	0.625	0.583	0.769	0.231	0.462

Table A.25: Participant7 J48 Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Work/School	0.5	0.667	0.5	0.5	0.5	0.667	0.667	0.5	0.833	0.5	0.333
Home	0.5	0.667	0.5	0.5	0.833	0.5	0.5	0.5	1	0.333	0.5
[GSU]Library	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.462	0.615	0.462	0.462	0.615	0.538	0.538	0.462	0.846	0.385	0.385

Table A.26: Participant7 RF Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Work/School	0.833	0.667	0.833	0.833	0.667	0.667	0.833	0.667	0.667	0.833	0.833
Home	0.667	0.833	0.667	0.667	0.667	0.667	0.5	0.667	1	0.333	0.5
[GSU]Library	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.692	0.692	0.692	0.692	0.615	0.615	0.615	0.615	0.769	0.538	0.615

Table A.27: Participant7 NB Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Work/School	0.5	0.667	0.3	0.5	0.883	0.667	0.667	0.667	1	0.5	0.667
Home	0.5	0.667	0.5	0.167	0.883	0.333	0.5	0.5	0.833	0.167	0.333
[GSU] Library	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.462	0.615	0.385	0.308	0.769	0.462	0.385	0.538	0.846	0.308	0.462

Table A.28: Participant7 KNN Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	1	0.913	0.826	0.87	0.875	0.739	0.739	0.696	0.957	0.87	0.87
On the Road,(Transportation)	0	0	0	0	0.3	0	0.250	0.250	0	0	0
Outdoor (Park, etc.)	0	0	0	0	0.667	0	0	0	0	0	0
Work/School	0	0	0	0.2	0.825	0.2	0.4	0.2	0	0	0
Weighted Average	0.639	0.583	0.528	0.581	0.583	0.5	0.556	0.5	0.611	0.556	0.556

Table A.29: Participant8 J48 Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	0.826	0.826	0.826	0.87	0.826	0.87	0.826	0.826	0.739	0.739	0.739
On the Road,(Transportation)	0.25	0	0	0	0.3	0	0.	0.250	0	0	0.25
Outdoor (Park, etc.)	0.25	0.5	0.75	0.5	0.667	0.5	0.75	0.5	0.5	0.5	0.5
Work/School	0	0.2	0	0	0	0	0	0	0	0	0
Weighted Average	0.611	0.611	0.611	0.611	0.611	0.583	0.611	0.583	0.528	0.528	0.556

Table A.30: Participant8 RF Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	0.739	0.826	0.826	0.609	0.739	0.696	0.783	0.783	0.957	0.696	0.652
On the Road,(Transportation)	0.25	0.5	0	0.5	0.25	0.5	0..25	0.250	0.25	0	0.25
Outdoor (Park, etc.)	0	0.5	0.75	0.75	0.5	0.5	0.5	0.5	0	0	0.25
Work/School	0.4	0.4	0	0	0	0.4	0.2	0.2	0	0	0.2
Weighted Average	0.556	0.694	0.611	0.528	0.556	0.611	0.611	0.583	0.639	0.444	0.5

Table A.31: Participant8 NB Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	0.739	0.739	0.652	0.739	0.696	0.783	0.696	0.739	0.739	0.696	0.652
On the Road,(Transportation)	0.25	0	0	0.5	0.5	0	0	0	0.	0	0.5
Outdoor (Park, etc.)	0.25	0.5	0.75	0.75	0.5	0.5	0.75	0.5	0.5	0.5	0.75
Work/School	0	0	0.2	0.2	0.2	0.2	0	0	0	0	0
Weighted Average	0.556	0.583	0.583	0.583	0.583	0.583	0.528	0.528	0.528	0.5	0.472

Table A.32: Participant8 KNN Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Work/School	0.563	0.688	0.438	0.719	0.75	0.781	0.656	0.719	0.813	0.344	0.031
Outdoor(Park, etc.)	0.286	0.714	0.286	0.657	0.771	0.657	0.657	0.743	0.743	0.686	0
On the Road (Transportation)	0.683	0.733	0.653	0.861	0.752	0.861	0.842	0.832	0.861	0.842	0.851
Mall/Shop	0	0	0	0	0.125	0.125	0	0	0	0	0
Home	0.618	0.794	0.529	0.618	0.765	0.735	0.559	0.765	0.735	0.5	0.412
Canteen/Restaurant/Cafe/Bar	0.276	0.621	0.172	0.517	0.759	0.69	0.448	0.759	0.862	0.103	0
Gym	0	0	0	0	0	0	0	0	0	0	0
[GSU] Canteen/garden	0	0.5	0.167	0.667	0.333	0.333	0.333	0.5	0.5	0	0
[GSU] Classroom/Lab	0	0	0	0	0	0.667	0	0.667	0	0	0
Other	0	0	0	0	0.571	0	0	0	0	0	0
Weighted Average	0.492	0.66	0.454	0.676	0.711	0.723	0.637	0.727	0.75	0.547	0.395

Table A.33: Participant9 J48 Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Work/School	0.375	0.75	0.531	0.563	0.813	0.719	0.563	0.781	0.781	0.344	0.25
Outdoor(Park, etc.)	0.286	0.8	0.429	0.8	0.714	0.857	0.771	0.857	0.714	0.657	0.229
On the Road (Transportation)	0.683	0.881	0.634	0.881	0.881	0.921	0.881	0.921	0.782	0.851	0.653
Mall/Shop	0	0.25	0.125	0.125	0.125	0.25	0.125	0.25	0	0	0
Home	0.559	0.794	0.735	0.676	0.853	0.824	0.735	0.824	0.676	0.471	0.412
Canteen/Restaurant/Cafe/Bar	0.379	0.793	0.483	0.655	0.759	0.793	0.448	0.759	0.724	0.138	0.138
Gym	0	0	0	0	0	0	0	0	0	0	0
[GSU] Canteen/garden	0	0.667	0.667	0.667	0.667	0.667	0.333	0.667	0.667	0.667	0.167
[GSU] Classroom/Lab	0.667	0.333	0	0	0	0	0	0	0	0	0
Other	0	0.429	0	0.143	0.286	0.143	0.286	0.143	0	0	0
Weighted Average	0.473	0.785	0.547	0.703	0.773	0.797	0.723	0.801	0.691	0.563	0.395

Table A.34: Participant9 RF Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Work/School	0.344	0.656	0.719	0.125	0.656	0.281	0.188	0.25	0.375	0	0.375
Outdoor(Park, etc.)	0.286	0.743	0.2	0.657	0.6	0.657	0.657	0.657	0.543	0.686	0
On the Road (Transportation)	0.515	0.723	0.347	0.723	0.713	0.772	0.693	0.772	0.733	0.762	0.248
Mall/Shop	0.125	0.5	0.25	0.125	0.375	0.375	0	0.375	0	0	0
Home	0.676	0.824	0.706	0.588	0.765	0.706	0.647	0.735	0.735	0.206	0.382
Canteen/Restaurant/Cafe/Bar	0.345	0.724	0.241	0.448	0.724	0.759	0.586	0.724	0.862	0.123	0
Gym	0	0	0	0	0	0	0	0	0	0	0
[GSU] Canteen/garden	0	0.167	0.5	1	0.667	1	1	1	0.333	0.14	0.333
[GSU] Classroom/Lab	0	0	0	0	0	0.333	0	0	0.333	0.103	0
Other	0	0	0	0	0.143	0	0	0	0	0	0
Weighted Average	0.418	0.68	0.395	0.547	0.66	0.648	0.563	0.641	0.617	0.086	0.203

Table A.35: Participant9 NB Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Work/School	0.469	0.594	0.438	0.531	0.5	0.563	0.5	0.469	0.781	0.438	0.313
Outdoor(Park, etc.)	0.286	0.743	0.457	0.743	0.6	0.8	0.771	0.771	0.743	0.6	0.286
On the Road (Transportation)	0.634	0.733	0.554	0.822	0.634	0.802	0.723	0.762	0.792	0.782	0.535
Mall/Shop	0	0.25	0	0.25	0.	0.25	0.125	0.125	0	0	0
Home	0.559	0.765	0.588	0.765	0.735	0.794	0.765	0.794	0.706	0.412	0.265
Canteen/Restaurant/Cafe/Bar	0.207	0.759	0.414	0.621	0.586	0.621	0.621	0.655	0.621	0.241	0.069
Gym	0	0	0	0	0	0	0	0	0	0	0
[GSU] Canteen/garden	0	0.667	0.667	0.667	0.667	0.667	0.667	0.667	1	0.667	0
[GSU] Classroom/Lab	0.333	0.333	0	0	0	0	0	0	0	0	0
Other	0	0.286	0.286	0.286	0	0.286	0.286	0.286	0	0.286	0
Weighted Average	0.439	0.688	0.484	0.695	0.695	0.652	0.652	0.672	0.699	0.551	0.332

Table A.36: Participant9 KNN Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
On the Road,(Transportation)	0.567	0.733	0.4	0.9	0.667	0.9	0.867	0.767	0.733	0.8	0.3
Work/School	0.3	0.3	0.3	0.5	0.3	0.6	0.6	0.4	0.1	0.1	0
Canteen/Restaurant/Cafe/Bar	0	0	0	0	0	0	0	0	0	0	0
Outdoor (Park, etc.)	0	0	0	0.5	0	0.333	0.333	0.5	0.167	0.667	0
Mall/Shop	0	0	0	0	0	0	0	0	0	0	0
Home	0.042	0.266	0.708	0.458	0.667	0.625	0.5	0.625	0.708	0.708	0.625
Other	0	0	0	0	0	0	0	0	0	0	0
[GSU] Classroom/Lab	0.4	0.2	0.6	0.2	0.2	0	0.8	0	0	0	0
Gym	0	0.25	0	0	0.5	0	0	0	0.75	0	0
Weighted Average	0.284	0.5	0.284	0.534	0.477	0.568	0.568	0.511	0.5	0.523	0.273

Table A.37: Participant10 J48 Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
On the Road,(Transportation)	0.333	0.733	0.267	0.967	0.333	0.933	0.967	0.933	0.867	0.867	0.2
Work/School	0.7	0.6	0.8	0.6	0.6	0.6	0.7	0.6	0.3	0.4	0.1
Canteen/Restaurant/Cafe/Bar	0.333	0.333	0	0.333	0	0	0	0	0	0	0
Outdoor (Park, etc.)	0.333	0.333	0.333	0.667	0.333	0.667	0.667	0.667	0.333	0.833	0.333
Mall/Shop	0	0.333	0.2	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0
Home	0	0.2	0.333	0.708	0.708	0.75	0.708	0.75	0.708	0.625	0.458
Other	0.292	0.266	0	0	0	0	0	0	0	0	0
[GSU] Classroom/Lab	0.6	0.6	0.4	0.6	0.6	0.6	0.6	0.6	0.2	0	0
Gym	0	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.75	0.5	0.5
Weighted Average	0.318	0.602	0.352	0.727	0.625	0.716	0.727	0.716	0.545	0.614	0.25

Table A.38: Participant10 RF Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
On the Road,(Transportation)	0.5	0.767	0.133	0.733	0.733	0.867	0.9	0.933	0.8	0.6	0.167
Work/School	0.5	0.4	0.5	0.2	0.3	0.2	0.3	0.6	0	0.1	0
Canteen/Restaurant/Cafe/Bar	0	0.333	0.333	0.333	0.333	0.333	0	0	0	0.333	0
Outdoor (Park, etc.)	0.833	0.5	0.333	0.667	0.333	0.667	0.5	0.667	0.333	0.5	0
Mall/Shop	0	0	0.2	0.2	0.2	0.2	0.4	0.4	0	0	0
Home	0.208	0.792	0.583	0.25	0.75	0.417	0.583	0.75	0.833	0.042	0.583
Other	0	0	0	0	0	0	0	0	0	0	0
[GSU] Classroom/Lab	0	0.4	0.6	0.6	0.4	0.6	0.6	0.6	0.2	0.6	0
Gym	0.5	0.5	0.5	0	0.5	0.5	0.5	0.5	0.5	0	0.5
Weighted Average	0.364	0.614	0.364	0.433	0.58	0.557	0.614	0.716	0.557	0.307	0.239

Table A.39: Participant10 NB Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
On the Road,(Transportation)	0.467	0.567	0.433	0.733	0.5	0.733	0.6	0.633	0.7	0.8	0.333
Work/School	0.6	0.8	0.6	0.7	0.6	0.6	0.6	0.6	0.3	0.4	0.2
Canteen/Restaurant/Cafe/Bar	0.333	0.333	0	0.667	0	0.667	0.667	0.667	0	0	0
Outdoor (Park, etc.)	0	0.333	0.333	0.667	0.333	0.667	0.333	0.333	0.333	0.833	0.333
Mall/Shop	0	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0
Home	0.167	0.542	0.333	0.2	0.625	0.625	0.625	0.583	0.667	0.667	0.167
Other	0	0	0	0	0	0	0	0	0	0	0
[GSU] Classroom/Lab	0.6	0.4	0.4	0.6	0.6	0.6	0.6	0.6	0	0	0
Gym	0	0.25	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Weighted Average	0.318	0.523	0.398	0.614	0.511	0.511	0.568	0.586	0.523	0.602	0.227

Table A.40: Participant10 KNN Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Other	0	0	0.333	0	0.333	0	0	0.333	0	0	0.333
Work/School	0.789	0.719	0.667	0.737	0.737	0.754	0.719	0.737	0.614	0.544	0.526
Home	0.717	0.733	0.733	0.75	0.8	0.783	0.8	0.85	0.8	0.45	0.6
[GSU] Classroom/Lab	0.829	0.61	0.634	0.659	0.634	0.61	0.659	0.683	0.195	0.122	0.317
[GSU] Canteen/garden	0	0.25	0.313	0.188	0.313	0.313	0.25	0.375	0.438	0.313	0.375
Outdoor (Park, etc.)	0.167	0.167	0.25	0.25	0.083	0.292	0.333	0.208	0.042	0.458	0.125
Canteen/Restaurant/Cafe/Bar	0	0	0	0	0	0	0	0	0	0	0
On the Road (Transportation)	0.24	0.76	0.36	0.8	0.76	0.72	0.72	0.6	0.6	0.68	0.36
Weighted Average	0.567	0.588	0.554	0.614	0.614	0.622	0.627	0.635	0.489	0.412	0.421

Table A.41: Participant11 J48 Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Other	0	0	0	0	0.667	0	0.333	0.667	0	0	0.333
Work/School	0.719	0.825	0.807	0.807	0.86	0.877	0.807	0.842	0.684	0.632	0.614
Home	0.717	0.8	0.817	0.733	0.867	0.8	0.8	0.85	0.7	0.55	0.55
[GSU] Classroom/Lab	0.731	0.659	0.756	0.732	0.707	0.732	078	0.756	0.195	0.195	0.293
[GSU] Canteen/garden	0.125	0.313	0.5	0.5	0.5	0.438	0.438	0.375	0.375	0.438	0.25
Outdoor (Park, etc.)	0.208	0.208	0.25	0.458	0.208	0.458	0.417	0.458	0.292	0.5	0.292
Canteen/Restaurant/Cafe/Bar	0	0	0	0.286	0	0	0.286	0	0	0	0
On the Road (Transportation)	0.4	0.52	0.48	0.48	0.56	0.68	0.56	0.64	0.44	0.72	0.2
Weighted Average	0.562	0.622	0.652	0.657	0.682	0.7	0.687	0.708	0.485	0.489	0.412

Table A.42: Participant11 RF Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Other	0	0	0.333	0.333	0.667	0.333	0.667	0.667	0	0.333	0.333
Work/School	0.772	0.789	0.702	0.491	0.772	0.456	0.439	0.526	0.579	0	0.368
Home	0.717	0.717	0.667	0.567	0.783	0.6	0.667	0.733	0.883	0.283	0.8
[GSU] Classroom/Lab	0.512	0.453	0.512	0.439	0.317	0.488	0.537	0.488	0.049	0.366	0.195
[GSU] Canteen/garden	0	0.5	0.188	0.313	0.25	0.25	0.313	0.438	0.438	0.188	0.125
Outdoor (Park, etc.)	0.042	0.083	0.208	0.5	0.167	0.5	0.5	0.458	0.208	0.458	0.208
Canteen/Restaurant/Cafe/Bar	0	0.143	0	0	0.143	0	0	0	0	0	0.143
On the Road (Transportation)	0	0.52	0	0.8	0.68	0.8	0.84	0.88	0.36	0.12	0.04
Weighted Average	0.468	0.562	0.481	0.506	0.567	0.519	0.545	0.584	0.468	0.288	0.103

Table A.43: Participant11 NB Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Other	0	0.667	0.333	0.667	0.333	0.667	0.667	0.667	0	0	0
Work/School	0.807	0.789	0.684	0.754	0.702	0.772	0.702	0.702	0.719	0.684	0.649
Home	0.75	0.75	0.817	0.733	0.85	0.733	0.833	0.833	0.7	0.567	0.5
[GSU] Classroom/Lab	0.756	0.61	0.610	0.732	0.561	0.659	0.659	0.585	0.146	0.268	0.341
[GSU] Canteen/garden	0	0.375	0.438	0.438	0.5	0.5	0.5	0.5	0.375	0.5	0.375
Outdoor (Park, etc.)	0.167	0.167	0.208	0.292	0.25	0.417	0.292	0.333	0.167	0.417	0.25
Canteen/Restaurant/Cafe/Bar	0	0	0.143	0.286	0.143	0.143	0.286	0.286	0	0	0
On the Road (Transportation)	0.36	0.48	0.24	0.44	0.28	0.44	0.44	0.44	0.32	0.52	0.077
Weighted Average	0.597	0.597	0.571	0.627	0.581	0.631	0.631	0.622	0.458	0.494	0.425

Table A.44: Participant11 KNN Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
On the Road,(Transportation)	0.636	0.545	0.727	0.909	0.727	0.909	0.909	0.909	0.636	0.818	0.818
Outdoor (Park, etc.)	0	0	0	0	0	0	0.333	0.333	0	0.684	0.649
Work/School	0.6	0	0	0	0.2	0	0	0	0.6	0.567	0.5
Home	0.833	0.667	0.667	0.667	0.667	0.667	0.667	0.667	0.833	0.268	0.341
[GSU] Canteen/garden	0	0	0	0	0	0	0	0	0	0	0
[GSU] Classroom/Lab	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.536	0.357	0.429	0.5	0.464	0.5	0.536	0.536	0.563	0.464	0.429

Table A.45: Participant12 J48 Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
On the Road,(Transportation)	0.455	0.364	0.727	0.727	0.818	0.818	0.818	0.909	0.455	0.909	0.818
Outdoor (Park, etc.)	0	0	0	0	0	0	0	0.333	0	0	0
Work/School	0.6	0.6	0.6	0.4	0.6	0.4	0.6	0.4	0.6	0.4	0
Home	0.5	0.667	0.833	0.667	0.667	0.667	0	0.833	0.333	0.5	0.5
[GSU] Canteen/garden	0	0	0	0	0	0	0	0	0	0	0
[GSU] Classroom/Lab	1	1	0.5	0.5	0.5	0.5	0.5	0	0	0	0
Weighted Average	0.464	0.464	0.607	0.536	0.607	0.571	0.607	0.607	0.357	0.536	0.429

Table A.46: Participant12 RF Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
On the Road,(Transportation)	0.237	0.455	0.727	0.636	0.909	0.727	0.727	0.909	0.455	0.636	0.818
Outdoor (Park, etc.)	0	0	0	0	0	0	0	0	0	0	0
Work/School	0.4	0.6	0.2	0.2	0.6	0.4	0.2	0.4	0.8	0	0
Home	0.5	0.5	0.333	0.833	0.667	0.833	0.667	0.667	0.667	0.667	0.167
[GSU] Canteen/garden	0	0	0	0	0	0	0	0	0	0	0
[GSU] Classroom/Lab	0	0	0	0	0.5	0	0	0	0	0	0
Weighted Average	0.286	0.393	0.393	0.464	0.607	0.536	0.464	0.571	0.464	0.393	0.357

Table A.47: Participant12 NB Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
On the Road,(Transportation)	0.455	0.454	0.454	0.727	0.818	0.727	0.727	0.727	0.636	0.909	0.818
Outdoor (Park, etc.)	0	0	0	0	0	0	0	0	0	0	0
Work/School	0.4	0.6	0.6	0.4	0.6	0.4	0.4	0.4	0.6	0.4	0.2
Home	0.333	0.833	0.667	0.833	0.667	0.833	0.5	0.833	0.333	0.333	0.333
[GSU] Canteen/garden	0	0	0	0	0	0	0	0	0	0	0
[GSU] Classroom/Lab	1	1	0.5	1	0.5	1	1	1	0	0	0
Weighted Average	0.393	0.536	0.464	0.607	0.607	0.607	0.536	0.607	0.429	0.5	0.429

Table A.48: Participant12 KNN Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	1	0.9	1	0.9	0.9	0.9	0.9	0.9	0.9	0.9	0.9
Work/School	0.6	1	0.6	1	1	1	0.6	1	1	1	0.6
Canteen/Restaurant/Cafe/Bar	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.813	0.875	0.813	0.75	0.875	0.875	0.75	0.875	0.875	0.875	0.75

Table A.49: Participant13 J48 Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	1	1	1	0.9	1	1	1	1	0.9	0.9	0.9
Work/School	0.8	1	1	1	1	1	1	1	1	0.8	1
Canteen/Restaurant/Cafe/Bar	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.875	0.938	0.938	0.875	0.938	0.938	0.938	0.939	0.875	0.813	0.875

Table A.50: Participant13 RF Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	1	0.9	0.8	0.9	0.8	0.9	0.9	1	0.9	0.9	0.8
Work/School	0.6	0.8	0.4	0.2	0.6	0.2	0.2	1	0.2	0	0.4
Canteen/Restaurant/Cafe/Bar	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.813	0.813	0.625	0.625	0.688	0.625	0.5	0.625	0.625	0.563	0.625

Table A.51: Participant13 NB Feature Combinations Results

Labels	Time	TC	TP	TA	TCP	TCA	TPA	ALL	Comm	Act	Phone
Home	1	1	1	1	1	1	1	1	0.9	0.9	0.9
Work/School	1	1	1	1	1	1	1	1	1	1	0.8
Canteen/Restaurant/Cafe/Bar	0	0	0	0	0	0	0	0	0	0	0
Weighted Average	0.938	0.938	0.938	0.938	0.938	0.938	0.938	0.939	0.875	0.875	0.813

Table A.52: Participant13 KNN Feature Combinations Results

BIOGRAPHICAL SKETCH

Selek Ceren Çelik was born on January 01, 1987 in Eskişhir, Turkey. After graduating from Eskişehir Kılıçoğlu High School in 2005, she began studying Computer Engineering at Anadolu University in 2005. After graduating from in 2010 she began studying Computer Engineering for Master degree at Galatasaray University in 2010. Since September 2015, she is working as a research assistant at İstanbul Bilgi University.

PUBLICATIONS

— S.C. Çelik and Ö. Durmaz İncel, "Semantic Place Prediction From Mobile Phone Sensors" In 2016 24th Signal Processing and Communication Application Conference(SIU). IEEE, 2016. p. 1021-1024.