THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

CLASSIFICATION OF STUDENTS' MOODS FROM THEIR MOBILITY DATA

Master's Thesis

MERVE KAYA

İSTANBUL, 2019

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES COMPUTER ENGINEERING

CLASSIFICATION OF STUDENTS’ MOOD FROM THEIR MOBILITY

Master’s Thesis

MERVE KAYA

Thesis Supervisor: ASSIST. PROF. DR. ECE GELAL SOYAK
Co-Advisor: ASSOC. PROF. DR. ÖZLEM DURMAZ İNCEL

İSTANBUL, 2019

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
COMPUTER ENGINEERING

Name of the thesis : Classification of Students' Mood from Their Mobility
Name/Last Name of the Student: Merve KAYA
Date of the Defense of Thesis : 20/08/2019

The thesis has been approved by the Graduate School of Natural and Applied Sciences.

Assist. Prof. Dr. Yücel Batu SALMAN
Graduate School Director

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Tarkan AYDIN
Program Coordinator

This is to certify that we have read this thesis and we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Science.

Examining Committee Members ________Signature____ _____

Thesis Supervisor

Asst. Prof. Dr. Ece GELAL SOYAK ----------------------------------

Thesis Co-Supervisor

Assoc. Prof. Dr. Özlem DURMAZ İNCEL ----------------------------------

Member

Asst. Prof. Dr. Cemal Okan ŞAKAR ----------------------------------

Member

Asst. Prof. Dr. Tuna ÇAKAR ----------------------------------

Member

Asst. Prof. Dr. Tevfik AYTEKİN ----------------------------------

ACKNOWLEDGMENTS

First of all, I would like to thank my thesis advisor, Asst. Prof. Dr. Ece GELAL SOYAK, and my thesis Co-Adviser, Assoc. Prof. Dr. Özlem DURMAZ İNCEL, who has allowed me to work on this thesis. I’m very grateful for their support, insight, and invaluable help during the preparation of this thesis.

I would like to thank my thesis committee, consisting of Asst. Prof. Dr. Cemal Okan ŞAKAR, Asst. Prof. Dr. Tevfik AYTEKİN, and Asst. Prof. Dr. Tuna ÇAKAR. Their feedback increased the quality of this thesis.

I would also like to thank the University of Bahcesehir, the Department of Computer Engineering and my professors for these years of teaching.

Last but not the least, I would like to thank my family: my parents, Nurten and Veli, my brother Mehmet Orhun for supporting me spiritually throughout writing this thesis and my life in general. They are my inspiration and motivation. Thank you for all your sacrifices and for always believing in me.

I also place on record, my sense of gratitude to everyone who directly or indirectly, have lent his or her hand in this venture.

İstanbul, 2019 Merve KAYA

ABSTRACT

CLASSIFICATION OF STUDENTS’ MOOD FROM THEIR MOBILITY

Merve KAYA

Computer Engineering Master Program

Thesis Supervisor: Asst. Prof. Dr. Ece GELAL SOYAK
Co-Advisor: Assoc. Prof. Dr. Özlem DURMAZ İNCEL

August 2019, 41 pages

Many factors affect how we feel, such as our daily experiences, weather, and exam periods for students. Sensing and wearable-mobile technologies make it possible to collect digital data about these factors to track our moods. The smartphones’ sensing abilities make them effective platforms for collecting different kinds of data such as social interactions, mobility and sleeping. In this respect, StudentLife (R.Wang et al., 2014) open dataset was collected to track students’ behaviors over the course of a 10-week academic year, from 48 students at Dartmouth University. In this thesis, our target is to investigate the relationship between students’ mood with their mobility patterns using the StudentLife dataset

In this thesis, the students’ mood (happiness and sadness levels), Piazza usage, sleeping and mobility parameters from the StudentLife dataset are used. The mood data is taken as the target variable to classify and for each mood data, its response time and student id are merged with the sleeping, piazza usage and current and daily average mobility variables for that timestamp. For the data analysis, the WEKA platform is used. Four main classifiers are used for analyzing the data: J48 (decision tree), Random Forest, SVM and MLP. Also, attribute selection methods and PCA are applied for exploring the effect on the accuracy. The effect of mobility is investigated by excluding/including the mobility-related attributes. Our results show that attribute selection with mobility variables reveal better results. The highest accuracy is achieved with the MLP classifier, at around 67.2131% with default parameters; and increases to 70-72% for higher training time. In addition, we have seen that attribute selection, when applied with both MLP and SVM algorithms, chooses location and mobility-related attributes to improve algorithm accuracy. Based on our results, we interpret that location and mobility patterns have an impact on mood data; but for more accurate results, the mood classification model of each student should be personalized and more data should be collected. This is mainly due to the fact that the mood data is not measurable and each person has different happiness parameters.

Keywords: Mood, Mobility, Mobile Sensing, Classification, Students

ÖZET

ÖĞRENCİLERİN AKTİFLİK DURUMLARINDAN, RUH HALLERİNİN SINIFLANDIRILMASI

Merve KAYA

Bilgisayar Mühendisliği Yüksek Lisans Programı

Tez Danışmanı: Dr. Öğr. Üyesi Ece GELAL SOYAK
Ek Danışman: Doç. Dr. Özlem DURMAZ İNCEL

Ağustos 2019, 41 sayfa

Ruhsal durumu birçok etkileyen durum olabilir. Akıllı telefon sensörleri pek çok bilgiyi toplayabilmemize olanak sağlıyor. Sosyal etkileşim, hareketlilik veya uyku gibi bilgileri bu sensörler ile toplayabiliriz. Bu durum 38 öğrencinin davranışlarının, 10 haftalık bir akademik yıl boyunca incelenmesine olanak sağlıyor. StudentLife R.Wang et al. (2014) veri seti Dartmouth Üniversitesinden, 48 ön lisans ve lisans öğrencinin verisini içermektedir. Bu çalışma ile öğrencilerin ruhsal durumunun onların hareketliliği ile olan ilişkisi incelenecektir.

Tez çalışmasında, StudentLife veri setinde bulunan ruhsal durum (mutluluk-üzüntü seviyeleri), piazza kullanımı, uyku verileri ile öğrencilerin günlük aktivite verilerden yararlanılmıştır. Ruhsal durum verisi günlük ve öğrenci numaraları bazında birleştirilmiştir. WEKA'da bulunan J48, Random Forest, SVM ve MLP gibi sınıflandırmalardan ve özellik seçimi ile PCA uygulamalarından yararlanılmıştır. Mobilete verisinin etkisini gözlemlemek için tüm uygulamalar mobilete verileri ile ve mobilete verileri elenerek analiz edilmiştir.

Özellik seçimi, mobilete verisi ile MLP sınıflandırmasında iyi sonuçlar vermiştir. En yüksek sonuç %67.2131 ile MLP algoritması üzerinden elde edilmiştir. MLP algoritması training sayıları arttırıldığında elde edilen sonuçlar %70-72 civarında olmuştur. PCA uygulaması mobilete verileri ile yapıldığında, sonuçları negatif olarak etkilenmiştir. PCA verileri lokasyon verileri bulunmadığında SVM sınıflandırması üzerinde daha iyi sonuçlar vermiştir. Özellik seçimi uygulanmadığında mobilete verileri ile çalıştırılan sınıflandırma algoritmalarının sonuçları daha başarılı olmuştur. Mobilete verilerinin öğrencilerin ruhsal durumları ile ilişki içinde olduğu gözlemlenmiştir. Ruhsal durumun kişiden kişiye değişmesinden dolayı, sonuçların da bu durumdan etkilendiği düşünülmektedir.

Anahtar Kelimeler: Ruh Hali, Hareketlilik, Mobil Sensör, Veri Madenciliği, Öğrenciler

CONTENTS

TABLES

FIGURES

ABBREVIATIONS

AP	:	Access Point
CSV	:	Comma Separated Values
EMA	:	Ecological Momentary Assessment
Freq.	:	Frequency
GPS	:	Global Positioning System
JSON	:	JavaScript Object Notation
ML	:	Machine Learning
MLP	:	Multilayer Perceptron
OUI	:	Organizationally Unique Identifier
PCA	:	Principal Component Analysis
SVM	:	Support Vector Machines
TPR	:	True Positive Rate
UNIX	:	Uniplexed Information and Computing Service
UTC	:	Coordinated Universal Time
WEKA	:	Waikato Environment for Knowledge Analysis

SYMBOLS

Sum of attributes	:	$\sum$
Reference time	:	t_0

1. INTRODUCTION

Smartphones can collect continuous sensor data, which may be analyzed to identify patterns over time. The recent literature contains research work on the analysis of such collected vitals, to reveal users' physical and behavioral patterns over different periods. For instance, A. Ghandeharioun et al. (2017) established a correlation between wearable or smartphone sensor data and users' psychology or habits.

To facilitate such an analysis, different open data sets are available for researchers. In this thesis, we investigate the relation between the students' measured data and their mood, on the StudentLife data set. This data set comprises students' daily answers to survey questions as well as input from the various sensors on the mobile phone. We analyze whether students' inferred mobility patterns have an impact on their reported mood ratings. To answer this research question, we examined the StudentLife public data set's self-reported mood reports, along with the sensor data indicating movement and location. Our aim is to build a classification model for mood prediction based on the mobility data along with other types of data, such as Piazza usage. For this purpose, we translate the WiFi statistics on the smartphone radio to location statistics; in addition, we also consider the measured GPS data indicating the student locations. We classify students' reported mood results based on HappyResult attribute. To examine how variability in mobility patterns are related to mood ratings, we introduce the impact of the LocationBeenCount, UniqueAPS, UniqueOUICount and each distinct location names that the students' have been, new features we have constructed using this analysis and evaluate the change in classification accuracy. We have used four different classifiers, namely, J48, Random Forest, SVM and MLP.

Our main contribution is the exploration of mobility relation on students' happiness level. While we achieved only 26.6917% accuracy when performing mood classification using MLP algorithm based on sleep and grades data, this accuracy increased to 67.2131% on attribute selection with mobility patterns. In addition, we observed that attribute selection algorithms (such as cfsSubSetEval, classifierSubSetEval) favored the mobility attributes to yield higher accuracies overall algorithms. These results led us to suspect that mobility patterns do influence the mood.

However, since the dataset is limited (i.e., many students have missing mood data), and the students' responses to survey questions regarding their moods may be subjective, we are unable to provide a conclusion indicating a certain correlation between mobility and location patterns and students moods.

The rest of the thesis is organized as follows. In Chapter 2, we explain related prior art. In Chapter 3 we provide an overview of the algorithms used in our study. In Chapter 4, we describe our data set and explain how the data is processed before the data mining algorithms are applied. In Chapter 5 we present the classification performances on the dataset with different algorithms and discuss our results; and Chapter 7 concludes the thesis with a discussion of future work.

2. RELATED WORK

In recent years, mobile sensing has been used as an unobtrusive method to track and model human behavior and health (M.S.H. Aunget, (2016), etc.). In this Chapter, we provide a brief summary of various prior studies that mostly observed that sleeping patterns over mental health, mobility patterns over physiological effects and depression-stress relation analysis with the mobility and sleeping patterns.

2.1 STUDIES FOCUSING ON BEHAVIORAL PATTERNS

Several prior studies investigated the relationship between mobile sensor data and personality traits. W. Wang et al. studied the variability in students' behavior using mobile sensing; they examined to what extent day-to-day behavioral patterns reveal a person's personality traits, on a data set collected over 646 students in the University of Texas Austin during two weeks. The result of predictions is reported to be correlated with self-reported personality traits. The results also show that smartphone data has the potential for passive personality assessment. Ben- Zeev et al. (2017) used mobile phones to collect passive sensing data from smartphones and on this data, they've examined a correlation between schizophrenia relapse signals and users' location, activity, and conversation statistics. In the research study, it is suggested creating innovative data management, modeling, and signal- detection techniques for being capable of these data to improve any treatment. In A. Ghandeharioun et al. (2017), the authors applied machine learning techniques on passively captured data via wearable wristbands to show correlation between mental health and sleep patterns. It is reported that poor mental health is accompanied by more irregular sleep, less motion, less incoming messages, less location patterns.

Wa L. et al (2013), used a broad array of built-in mobile phone sensors to predict mood, finding that decreases in calls, SMS messaging, Bluetooth-detected contacts, and location entropy were strongly related to feeling sad and stressed among students. C. David et al.(2017), is conducted using personal smartphones and wearable sensors to unobtrusively sense mental health state. As a result, found that personal sensing is still in its infancy, it holds great promise as a method for conducting mental health research.

All of these prior efforts focused on mental health from the perspective of depression or sociability. Our work differentiates itself from these efforts we mainly focused mobility patterns which covers WiFi and Wifi Location data and also for mood prediction used a new generated happyresult feature for making classification. In this thesis conversation activities of students are not in the scope.

2.2 PRIOR STUDIES ON THE STUDENTLIFE DATA SET

Several studies performed analysis on the StudentLife data set. In addition to mobile phone sensor data, this data set also relies on students' responses to surveys, providing data on students' perceived stress and happiness levels, sleep history, EMA answers for students' mood about how happy or stressful they are feeling at the time, and activities history.

Saeb et al. (2016) showed a correlation between students' mobility patterns plus phone usage, and students' depression levels by an analysis on the StudentLife data set. The research result shows that PHQ-9 scores are significantly correlated with the GPS features, location variance, entropy, and circadian movement. Gjoreski et al., (2015) investigated stress levels in students using the data from accelerometer, audio recorder, GPS, Wi-Fi, call log and light sensor of the smartphones. They resolved that the perceived stress is highly subjective and that only person-specific models are substantially better than the baseline.

Harari et al. (2017) analyzed the changes in students' activity and sociability behaviors over a term via the accelerometer and microphone sensors. They also worked on the StudentLife dataset and their results indicate that the students' activity levels and sociability have good levels (mean=1.87: activity, m=4.99: sociability) at the beginning of the terms. (M denotes the mean duration of the behavior for each week in hours). Also at the second term of the university, the activity level did not change however students' sociability increased. This study also worked for ethnic differences between students.

Boukhechba M. et al., (2017) is studied how non-invasive mobile sensing technology can be used to passively assess and predict social anxiety among college students. Over the GPS location, text messages, and call data collected from 54 college students over a two-week period indicates that social anxiety level can be predicted with an accuracy of up to 85%. The results are showed that mobility and communication patterns can help to reveal how social anxiety symptoms manifest in the daily lives of college students,

There is also prior work that investigated whether sleep patterns affect how we feel. Chen Z. et al., (2013) carried out an observation of over eight persons during one week. Their results are taken over two models BES model and a sleep-with-the-phone approach model which are the two popular commercial wearable systems. They found that the BES model's results are more successful.

Table 2.1: A Summary of Relevant Prior Work on StudentLife Data set

Reference	**Study**	**Findings**
S. Saeb et al.	Impact of mobility patterns on depression levels	The GPS features may be an important and reliable predictor of depressive symptom severity.
M. Gjoreski et al.	Classification of stress levels based on smartphone sensors	The perceived stress is highly subjective only person-specific models can be better than the baseline.
D. Ben-Zeev et al.	Integrating behavioral sensing and smartphone use for identifying digital indicators of psychotic relapse.	If data management will create innovatively, modeling, and signal-detection techniques, we can be capable of these data to improve any treatment.
A.Ghandeharioun	Objective assessment of depressive symptoms with machine learning	The poor mental health is highly correlated with irregular sleep,less motion, fewer incoming messages and less variability in location patterns.
G.M. Harari et al	Assessment of mental health, academic performance and behavioral	The first-term students' conversation level is lesser than the second term. The students start the term positively and

	trends of college students using smartphones	low stress. The term process all these inputs are increasing.
Boukhechba M. et al.	Monitoring Social Anxiety from Mobility and Communication Patterns	Over the GPS location, text messages, and call data found that social anxiety level can be predicted with high accuracies.
Chen Z. et al.	Unobtrusive Sleep Monitoring using Smartphones	Over the two different models, dedicated that BES model has better result of commercial wearable systems.

3. BACKGROUND

In this chapter, we briefly explain the classification algorithms that are used in this thesis. The research question that we explore is a classification problem and we use different classification algorithms to explore the performance. Before we explain the classification algorithms, it is shortly explained why the research question is a classification problem.

Classification is a data analysis task, which tries to find a model for describing and distinguishing data classes and concepts. It is about identifying a set of categories, which is the basis of a training set of data. The StudentLife dataset has happy/sad attribute, which is making the problem analysis based on the classification model to detect the other attributes which affect the mood data.

As the feature selection and classification algorithms, such as Feature Selection, J48, Random Forest, MLP and SVM are used and each of them is explained in the following.

3.1 FEATURE SELECTION

In data analysis, when there is a big data set with many attributes, it is challenging to identify which features should be used to create a predictive model. It may require good domain knowledge about the problem. The detection of the most relevant connected attributes with the selected feature is called “Feature Selection”. It is also known as variable or attribute selection. Mostly the feature selection is confused with the dimensionality reduction method. However, feature selection is not reducing the number of attributes over the dataset, while the dimensionality reduction method is creating new combinations of attributes.

3.2 PRINCIPAL COMPONENT ANALYSIS (PCA)

Principal Component Analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets. It is transforming the large set of variables into smaller ones.

The Principal Component Analysis steps are explained Lindsay I. Smith (2002):
Step 1: Standardization, on this step, the range of continuous initial variables are standardized. So each of them contributes equally to the analysis. Because if there will be a large difference of initial variables' ranges, those variables with larger ranges will dominate the others. Standardization formula mathematically is figured as Equation 1.

$$z = \frac{value - mean}{standard\ deviation}$$

Equation **1:** *Standardization formula mathematically*

The standardization will make all data on the same scale.

Step 2: Covariance Matrix computation,
On this step, the aim is to understand the variables of the input data set how they are varying from the mean and each other's which will make easier to understand the relationship of each other's. This step is making possible to decrease the effect of redundant variables over the data. This matrix is PxP symmetric matrix. (P is showing the number of dimensions). The creation of the matrix, the variance of each initial variables are taken. If sign of covariance is positive, the two variables increase or decrease together (correlated). If it is negative, one increases when the other decreases (inversely correlated).
Step 3: Compute the eigenvalues of the covariance matrix, the eigenvalues should be calculated from the covariance matrix for determining the principal component of the data. The principal components are new variables that are constructed as linear combinations or mixtures of the initial variables. For this purpose, if there will be 5 principal components, the data will give 5 principal components. Jaadi,Z..(2018) describe

the PCA is generally putting maximum possible information in the first component, then maximum remaining information to the second information and so on. The other thing that the principal components are interpretable and simply will not have any meaning since they combined linearly by initial variables.

Step 4: Feature Vector,

The eigenvalue computing and ordering as descending, allow to find the principal components significantly. In this step, to choose whether to keep all these components or discard those of lesser significance, we call the Feature vector. So, the feature vector is simply a matrix that has as columns the eigenvectors of the components that we decide to keep. This makes it the first step towards dimensionality reduction, because if we choose to keep only P eigenvectors (components) out of n, the final data set will have only p dimensions.

Last Step,

The aim of the last step is to use the feature vector formed using the eigenvectors of the covariance matrix, to reorient the data from the original axes to the ones represented by the principal components.

3.3 J48 DESICION TREE

The decision tree is a classic supervised learning algorithm. The decision tree learning is starting from the training data; create a predictive model, which is mapped to a tree structure. The goal of the algorithm is to achieve having good classification performance with a minimal number of decisions.

This algorithm generates the rules for the prediction of the target variables.

The basic algorithm of the decision tree:

1) start to take whole training dataset
2) over the best split, select attributes or values along dimension
3) based on each split' creates child nodes'
4) until the reach of stopping criteria, repeat all split to create new nodes. (stops when all examples have same class or the size of tree)

In this thesis when applying the J48 classifier, the default parameters of WEKA is used which is named as J48 –C 0.25 –M 2.The C4.5 algorithm for building decision trees is implemented in WEKA as a classifier called J48.
–C is the Confidence value (default 25%): lower values incur heavier pruning and -M is the Minimum number of instances in the two most popular branches (default)

3.4 RANDOM FOREST

Random forest, as its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.Yiu,T,(2018)

The random forest algorithm is summarized as follows:
Draw ntree samples from original data
From each sample, grow an unpruned classification or regression tree. At each node, randomly sampled mtry of the predictors and choose the best split from these variables.
Predict new data by aggregating the predictions of the ntree trees.

In the thesis Random forest classifier used for 100 and 500 iterations. The other Parameters of the classifier are used as WEKA's default parameters. The default parameters as set P-100 I-100 num-slots K-0 M 1.0 -V 0.001 -S 1

-**P** Size of each bag, as a percentage of the training set size. (Default 100)
-**I** <num> Number of iterations. (Current value 100)
-**num-slots** <num> Number of execution slots. (Default 1 - i.e. no parallelism)
(Use 0 to auto-detect number of cores)
-**K** <number of attributes> Number of attributes to randomly investigate.
(Default 0) (<1 = int(log_2(#predictors)+1))
-**M** <minimum number of instances> Set minimum number of instances per leaf. (default 1)

-V <minimum variance for split> Set minimum numeric class variance proportion
of train variance for split (default 1e-3).
-S <num> Seed for random number generator. (Default 1)

3.5 MULTILAYER PERCEPTRON (MLP)

A classifier uses backpropagation to learn a multi-layer perceptron to classify instances. The network parameters can also be monitored and modified during training time. The nodes in this network are all sigmoid (except for when the class is numeric, in which case the output nodes not become threshold linear units)

In this thesis, when applying the MLP algorithm, the default parameters are used. The default parameters as set -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 h a.

-L <learning rate> Learning rate for the backpropagation algorithm. (Value should be between 0 - 1, Default = 0.3).
-**M** <momentum> Momentum rate for the backpropagation algorithm. (Value should be between 0 - 1, Default = 0.2).
-**N** <number of epochs> Number of epochs to train through. (Default = 500).
-V <percentage size of validation set> Percentage size of validation set to use to terminate training (if this is non zero it can pre-empt num of epochs. (Value should be between 0 - 100, Default = 0).
-S <seed> the value used to seed the random number generator (Value should be >= 0 and and a long, Default = 0).
-E <threshold for number of consecutive errors> the number of consecutive increases of error allowed for validation testing before training terminates. (Value should be > 0, Default = 20).
-H <comma separated numbers for nodes on each layer>
The hidden layers to be created for the network. (Value should be a list of comma separated Natural numbers or the letters 'a' = (attribs + classes) / 2, 'i' = attribs, 'o' = classes,'t' = attribs. + Classes) for wildcard values, Default = a).

3.6 SUPPORT VECTOR MACHINES (SVM)

K.B.Lipkowitz(2007) et all explained the LIBSVM (Library for Support Vector Machines) that is developed by Chang and Lin and contains C-classification, ν-classification, ε-regression, and ν-regression. It is developed at C++ and Java and it supports multi-class classification, weighted SVM for unbalanced data, cross-validation and automatic model selection.

The SVM algorithm applies two different kernels, namely the Radial Basis Kernel and Sigmoid Kernel. In the following, we briefly explain these two algorithms. The kernel functions return the inner product between two points in a suitable feature space. Thus by defining a notion of similarity, with little computational cost even in very high-dimensional spaces.

3.6.1 Radial Basis Kernel

The kernel functions are used to map the original dataset (linear/nonlinear) into a higher dimensional space to make it a linear dataset. In machine learning, the **radial basis function** kernel, or **RBF** kernel is a popular **kernel function** used in various kernelized learning algorithms. It is generally using because it has localized and finite response along the entire x-axis.

3.6.2 Sigmoid Kernel

Hsuan-Tien Lin (2003), the sigmoid kernel was quite popular for support vector machines due to its origin from neural networks.

3.7 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, (M. A. Hall., 1998)

3.8 CLASSIFIERSUBSETEVAL

Evaluates attribute subsets on training data or a separate hold out testing set. Uses a classifier to estimate the 'merit' of a set of attributes.

4. DATA ANALYSIS AND MODELING

4.1 DESCRIPTION OF THE DATA SET

The StudentLife dataset is collected from 48 Dartmouth College students over a 10-week term. In addition to the data from various smartphone sensors, the data set is augmented with (i) students' responses to questionnaires on mental health outcomes (e.g., stress, happiness, loneliness, etc.), and (ii) students' activity in the online course platform Piazza.

The data is collected via a smartphone application which is named as StudentLife. The application senses the human behaviors. The application is working 7/24 hours on backstage of the smartphone that is making easier to collect all the data even the students are sleeping.

The dataset is anonymized by the provider so the students’ identities are hidden. That is why on the results students are declared as numbers. The students’ locations that they have been uniquely is counted and labeled by names. There are a hundred different location’s labels on the dataset. The detailed columns of names on the dataset are explained on the Table 4.11. In the following subsections, we explain the contents of the data and which parts are utilized in this Thesis.

4.1.1 Sensor Data

Sensor data set contains 10 different sensor data: physical activity, audio inferences, conversation inferences, Bluetooth scan, light sensor, GPS, phone charge, phone lock, WiFi, WiFi location. In this thesis we have only used GPS, WiFi, and WiFi Location data.

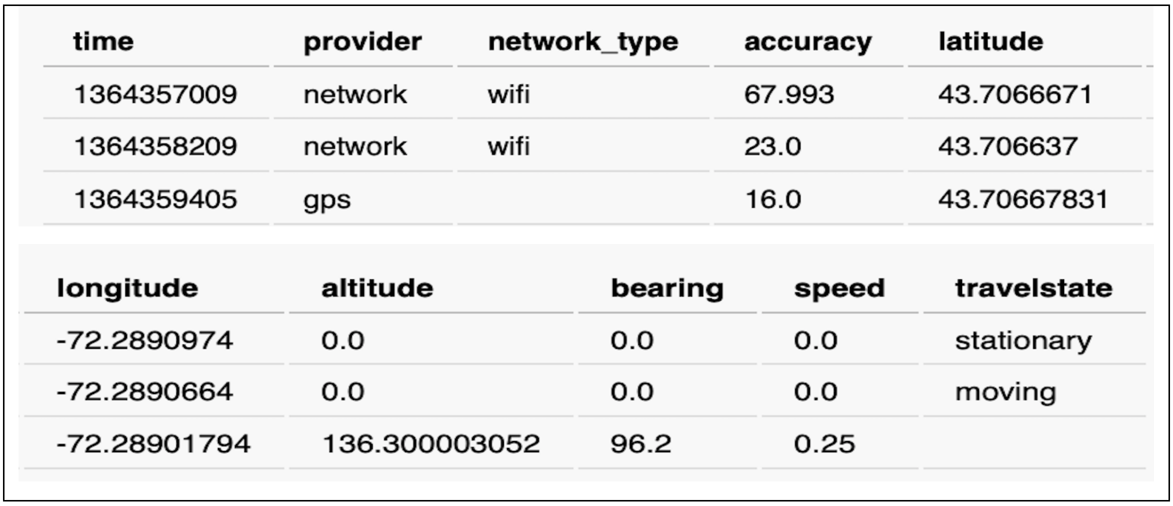
4.1.1.1 Gps Data

The GPS location data has time, provider, network-type, accuracy, latitude, longitude, altitude, bearing, and speed and travel state information, as shown in Table 4.1, and example data is shown in Figure 4.1. In this data, “travel state” field can be used for indicating students’ mobility. Although there are more columns, only the time and travelstate columns are used for merging data with the mood dataset. However on the dataset there were hundreds of missing values which were almost half of the data. That’s why it did not give the efficiency over the accuracy much that we were expecting.

Table 4.1: Description of Fields in the GPS Data

Field Name	Descriptions
time	time value in UNIX time format
provider	two types of providers, network or GPS
network_type	When the provider is network, it is specifying which network is used to obtain location.
travelstate	It is the information of movement or stationary situation according to the tagged GPS data.

Figure 4.1 Exemplary GPS Location Data View



time	provider	network_type	accuracy	latitude	longitude	altitude	bearing	speed	travelstate
1364357009	network	wifi	67.993	43.7066671	-72.2890974	0.0	0.0	0.0	stationary
1364358209	network	wifi	23.0	43.706637	-72.2890664	0.0	0.0	0.0	moving
1364359405	gps		16.0	43.70667831	-72.28901794	136.300003052	96.2	0.25	

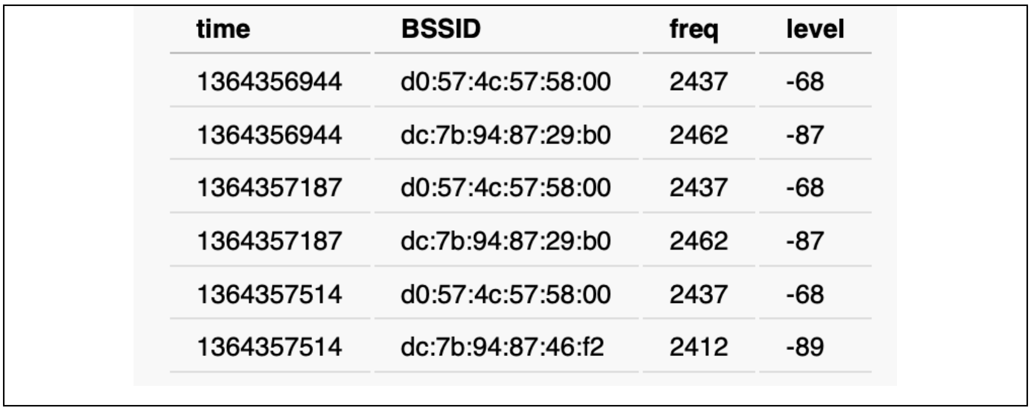
4.1.1.2 Wifi Data

Wi-Fi data includes four main columns: *Time*, *BSSID*, *frequency* and *level* (field definitions are described in Table 4.2, and sample values are shown in Figure 4.2).
Since our research problem is highly concerned with daily movements, and the frequency of the Access Point (AP) that the student connects to is not relevant to this detection, we have omitted the *freq* field. *level* field indicates the Received Signal Strength Indicator (RSSI) measured by the smartphone from the signals received from the AP that it is connected to. Thus, *level* can in fact indicate location and hint movement. However, WiFi signal strength attenuates rapidly even for the same location (Rappaport, T. S. (2002)); consequently, we could not use it as an accurate indicator of location. As a result, in this thesis, we have only used the *BSSID* field from the WiFi data set.

Table 4.2: Description of Fields in the WiFi Data

Field Name	Descriptions
time	time value which is UNIX time format
BSSID	WiFi Access Point(AP) MAC address
freq	Access Point channel frequency
level	WiFi signal strength

Figure 4.2: Exemplary WiFi Data View



time	BSSID	freq	level
1364356944	d0:57:4c:57:58:00	2437	-68
1364356944	dc:7b:94:87:29:b0	2462	-87
1364357187	d0:57:4c:57:58:00	2437	-68
1364357187	dc:7b:94:87:29:b0	2462	-87
1364357514	d0:57:4c:57:58:00	2437	-68
1364357514	dc:7b:94:87:46:f2	2412	-89

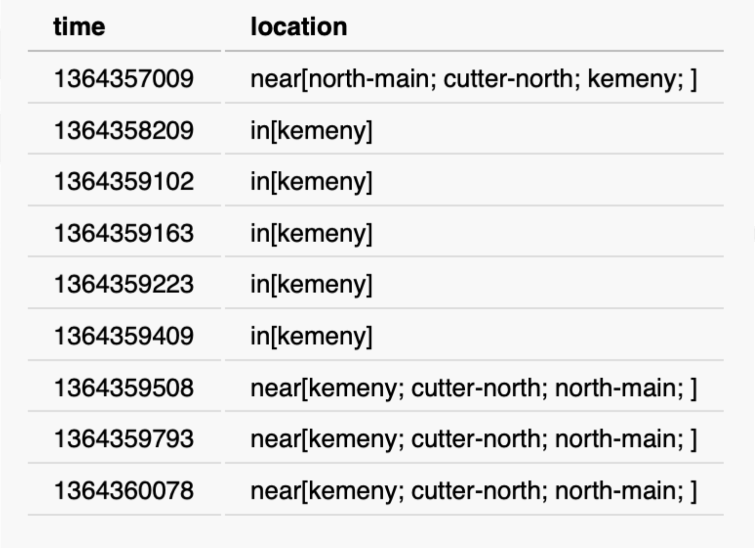
4.1.1.3 Wifi Location Data

WiFi location data has mainly two columns: time and location (Table 4.3, Figure 4.3). On the WiFi location data, two types of locations are inferred: in a building (e.g. in [in [kemeny]) and near some buildings (near[near [kemeny; cutter-north; north-main;main ;]).

Table 4.3: Description of Fields in the WiFi Location Data

Field Name	Descriptions
time	time value which is UNIX time format
location	Campus location information inferred by WiFi locations.

Figure 4.3: Exemplary WiFi Location Data View



time	location
1364357009	near[north-main; cutter-north; kemeny;]
1364358209	in[kemeny]
1364359102	in[kemeny]
1364359163	in[kemeny]
1364359223	in[kemeny]
1364359409	in[kemeny]
1364359508	near[kemeny; cutter-north; north-main;]
1364359793	near[kemeny; cutter-north; north-main;]
1364360078	near[kemeny; cutter-north; north-main;]

4.1.2 Ecological Momentary Assessment (Ema) Da

Wang,R (2014) defined the meaning of EMA as StudentLife dataset which is component to probe students’ states (e.g., stress, mood) across the term. EMA data is collected by the application StudentLife through a questionnaire asking multiple-choice questions about students’ daily mood states, stress levels and sleeping hours. EMA data includes different data labels however in this research the **sleeping**, **mood**, **mood2 (stress-related)** data are utilized and merged.

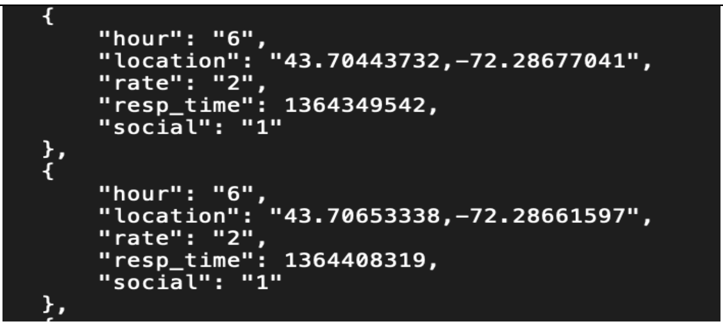
4.1.2.1 Sleeping Data

The overlook of the Sleeping questions and answers is given in Figure 4.4 and Table 4.4. The sleeping EMA questions are covering daily sleeping hour data, the rating of the sleeping quality and how much hard for them for staying awake during the day

Table 4.4: Sleeping Data EMA Questions

hour	How many hours did you sleep last night?	[1]<3 --- [2]3.5, [3]4 --- [4]4.5, [5]5 --- [6]5.5, [7]6 --- [8]6.5, [9]7 --- [10]7.5, [11]8---[12]8.5, [13]9---[14]9.5, [15]10---[16]10.5, [17]11---[18]11.5--[19]12,
rate	How would rate your overall sleep last night?	[1]Very-good, [2]Fairly-good, [3]Fairly-bad, [4]Very bad
social	How often did you have trouble staying awake yesterday while in class, eating meals or engaging in social activity?	[1]None, [2]Once, [3]Twice, [4]Three or more times

Figure 4.4: Sleeping Questions' Answer View



```
{
    "hour": "6",
    "location": "43.70443732,-72.28677041",
    "rate": "2",
    "resp_time": 1364349542,
    "social": "1"
},
{
    "hour": "6",
    "location": "43.70653338,-72.28661597",
    "rate": "2",
    "resp_time": 1364408319,
    "social": "1"
},
```

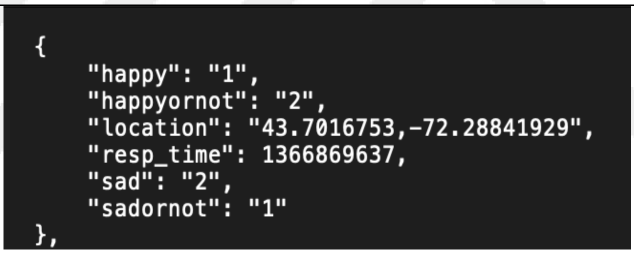
4.1.2.2 Mood Data

Mood data includes answers to the questions on whether the students are happy or sad, and how happy/sad they are. The structure of these questions are shown in Table 4.5 and a sample answer is shown in Figure 4.5.

Table 4.5: Mood Data EMA Questions

happyornot	Do you feel AT ALL happy right now?	(Yes) 1 (No) 2
happy	If you answered \"Yes\" on the first question, how happy do you feel?	[1]a little bit, [2]somewhat, [3]very much, [4]extremely,
sadornot	Do you feel AT ALL sad right now?	(Yes) 1 (No) 2
sad	If you answered \"Yes\" on the third question, how sad do you feel?	[1]a little bit, [2]somewhat, [3]very much, [4]extremely

Figure 4.5: Mood Questions' Answer View



```
{
    "happy": "1",
    "happyornot": "2",
    "location": "43.7016753,-72.28841929",
    "resp_time": 1366869637,
    "sad": "2",
    "sadornot": "1"
},
```

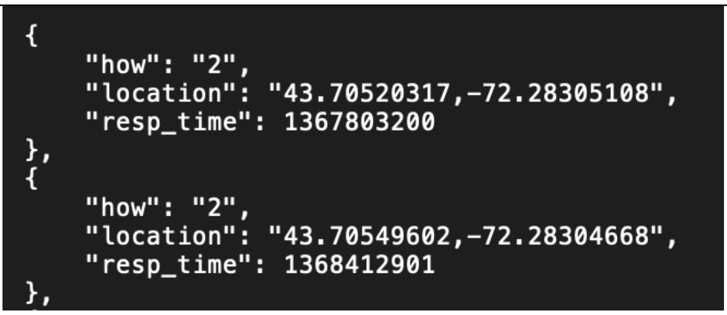
4.1.2.3 Mood2 (Stress-Related) Data

Mood2 data in the StudentLife data set is targeted to identify how students are feeling at the moment they are answering the questions on the application. The structure of these questions are shown in Table 4.6 and sample answer is shown in Figure 4.6. There are three degrees; happy, stressed and tired.

Table 4.6: Mood2 Data EMA Questions

how	How are you right now?	[1]happy [2]stressed [3]tired

Figure 4.6: Mood2 Questions' Answer View



```
{
    "how": "2",
    "location": "43.70520317,-72.28305108",
    "resp_time": 1367803200
},
{
    "how": "2",
    "location": "43.70549602,-72.28304668",
    "resp_time": 1368412901
},
```

4.1.3 Educational Data

Educational data contains; *(i)* the record of all students' academic performances, *i.e.* their **grades**, and *(ii)* their **Piazza** application usage statistics. Grade data contains, for each student, their cumulative average of their GPA as well as their grades indicating their academic performances in two courses at Dartmouth University.

4.1.3.1 Piazza Usage Data

The data is given mainly information based on application usage from Piazza, which is a platform via which Dartmouth University students can access course material and participate in online forums. Piazza data is described in Table 4.7.

Table 4.7: Description of Fields in Piazza Usage Data

Field Name	Description
Days Online	The number of days the student logged in CS65 Piazza class page
Views	Number of posts the students viewed
Contributions	Number of posts, responses, edits... On resume the anything at Piazza the student can follow up.
Questions	Number of questions students has asked.
Notes	Numbers of notes students has posted.
Answers	Numbers of questions students answered.

4.2 DATA PREPROCESSING AND FEATURE ENGINEERING

In StudentLife dataset, some part of the data is presented in JSON format, such as EMA answers. They are converted to the .csv format for each student by using a Python script.

In the StudentLife dataset, some new attributes are created such as UniqueAPs, UniqueOUICount, LocationBeenCount, happyresult, happy_desc, sad_desc and happysadlevel.

The first two features, UniqueAPs and UniqueOUICount, are created over the WIFI dataset described in Section 4.1.1.2 and Section 4.2.2.

The classification of StudentLife is done on mood (happy/sad) dataset using three different algorithms. These algorithms working processes are explained at Section 4.2.4.

After all processes, the data has 276 rows for 122 different attributes and 37 students.

4.2.1 Preprocessing Timestamps

Each value is converted as a column based on student ids. The entire student mobility variable is merged as a single file and merging is applied on the Database after all, the variables inserted. The main data is taken as Mood data. The STUDENTID and TIME variables are the unique key of the each rows. The time variable converted to the UTC format. (Figure 4.7)

Figure 4.7: SQL query used for retrieving to the time format UNIX to UTC

```
TO_DATE((SELECT to_char((TIMESTAMP '1970-01-01 00:00:00' +
numtodsinterval(W.RESPONSETIME, 'SECOND')), 'DD.MM.YYYY HH24:MI:SS')
ts FROM dual),'DD.MM.YYYY HH24:MI:SS')
```

4.2.2 Proprocessing Wifi Data

In addition, the WiFi data set is preprocessed, in order to identify the unique BSSIDs that students connect to each day.

We used the sum of the number of daily various different BSSIDs, which is tagged as uniqueAPs. The daily unique access point numbers are counted for detecting the student's mobility daily.

In addition, we tried to quantify how many different organizations the students have travelled to, daily. We have derived this data from the UniqueAPs data as follows: Organizationally Unique Identifier (OUI) represents the first 24 bits of the MAC address (the BSSID), which corresponds to the vendor of the WiFi Access Point whose beacons are detected. By counting the distinct OUIs instead of the distinct BSSIDs, we were able to distinguish whether the student moved across the coverage area of different APs inside a building, or whether the student has moved from a campus building to cafeteria area.

4.2.3 Preprocessing Wifi Location Data

In addition, we have also processed the WiFi Location data to a hundred (100) different location places are counted and stored as a new feature, which is named as LocationBeenCount.

Over the data for each location according to the studentid and daily duration been count is calculated and added as the new attributes. In this process, we aim to label how many times a student has been tagged by each of these place names.

4.2.4 Preprocessing Mood Data

In the thesis for the happy/sad assignment three different mood algorithms are used.

The first algorithm merges the “happyornot” and “happy” values into one column (“*HappyDesc*”), and “sadornot” and “sad” values into one column (“*SadDesc*”) and classifies the data over these values. According to Algorithm-1, *happyornot* and *happy* levels are merged such as 1.2, 1.3...etc., where the integer part belongs to the *happyornot* value and the fractional part belongs to the values are describing how happy that student reported at that time. When we classified all the dataset for happy and sad the final view of classified list for HappyDesc was {1.1, 1.2, 1.3, 1.4, and 2.1} and for the SadDesc {1.1, 1.2, 1.3, 1.4, and 2.1}.The pseudocode is shown in Table 4.8.

Table 4.8: Chart demonstrating the mapping of happy and sad mood data to a new field

```
if happyornot ==1 and happy == 1 then HappyDesc = 1.1
else if happyornot ==1 and happy == 2 then HappyDesc = 1.2
else if happyornot ==1 and happy == 3 then HappyDesc = 1.3
else if happyornot ==1 and happy == 4 then HappyDesc = 1.4
else if happyornot ==2 and happy == 1 then HappyDesc = 2.1
else if sadornot ==1 and sad == 1 then SadDesc = 1.1
else if sadornot ==1 and sad == 2 then SadDesc = 1.2
else if sadornot ==1 and sad == 3 then SadDesc = 1.3
else if sadornot ==1 and sad == 4 then SadDesc = 1.4
else if sadornot ==2 and sad == 1 then SadDesc = 2.1
```

As our initial results with algorithm-1 yielded very poor accuracy, we have designed new ways of mapping the mood data for classification. The second algorithm merges the “happy” and “sad” results into a single column named *HAPPYRESULT*. There are three main happiness levels. “*1*” indicates *happy*, “*2*”indicates *sad*. Some of the students did not specify their moods; thus, we have assigned the value “*3*” to such entries, to indicate “*neither happy nor sad*”. The pseudocode is shown in Table 4.10. When the mapping was complete, the rows of missing values, e.g. with *HAPPYRESULT* values of 0.1, 0.0 and 1.0 were eliminated from the data.

Table 4.9: Chart demonstrating of mapping mood data happy, sad, neither happy nor sad

```
if happyornot == 1 and sadornot==2 then  HAPPYRESULT = 1
if happyornot != 1 and sadornot == 1 then HAPPYRESULT =  2
if happyornot == 1 and sadornot == 1 and happy>sad then HAPPYRESULT= 1
if happyornot == 1 and sadornot == 1 and sad > happy then HAPPYRESULT =  2
if happyornot == 1 and sadornot == 1 and sad= happy then HAPPYRESULT= 3
```

The third algorithm is also similar to Algorithm-2, with a minor difference that is aimed at distinguishing the actual happiness levels of the students. Accordingly, we have introduced a variable for determining the difference between sad values. After attached the sad values as one for thirteen, two for twenty-three and three for thirty-three, four for forty-three. They are labelled as five, six, seven and eight. The happy range values did not modified that is why they are just taken as their happy values 1 to 4.

Table 4.10: Chart Demonstrating Our Mapping of Mood Range of Sad Data

```
IF HAPPYORNOT 1 HAPPY 1 THEN HAPPYSADLEVEL 1
IF HAPPYORNOT 1 HAPPY 2 THEN HAPPYSADLEVEL 2
IF HAPPYORNOT 1 HAPPY 3 THEN HAPPYSADLEVEL 3
IF HAPPYORNOT 1 HAPPY 4 THEN HAPPYSADLEVEL 4
IF SADORNOT 1 SAD 1 THEN HAPPYSADLEVEL 5
IF SADORNOT 1 SAD 2 THEN HAPPYSADLEVEL 6
IF SADORNOT 1 SAD 3 THEN HAPPYSADLEVEL 7
IF SADORNOT 1 SAD 4 THEN HAPPYSADLEVEL 8
```

After applying this algorithm to the dataset, we had an attribute, which is labeled as HAPPYSADLEVEL and the range for happy were between 1 to 4 and range for sad 5 to 8. In this algorithm also did not make the expected result for accuracy with the mobility.

We have observed that the second algorithm gave the most accurate results compared with the other two algorithms.

4.2.5 Preprocessing Educational Data (Grades)

All the students' results were not set meanwhile it was hard to make a connection between mood and GPA results. Some missing GPA results are assigned as E if the students are using Piazza regularly and if they are attending classes. After assigning the GPA results A to E according to the Dartmouth University alphabetic GPA board. Cause Of the missing values there were not directly connection with mood data.

Summary: Before making preprocessing normally we had five more attributes, namely *happysadlevel* (3th Algorithm), *happydesc* (2nd Algorithm), *saddesc* (2nd Algorithm), *stationarycount* (from GPS data) and *movingcount* (from GPS data). The *happysadlevel*, *saddesc* and *happdesc* are eliminated over the data because their accuracy is not giving satified results. Also the stationarycount and movingcount values that collected over GPS data, were not much supportive for relation between mood and mobility. After preprocessing is complete, our data has 122 different attributes (Table 4.11). After the data is ready, the open source machine learning tool, WEKA[1] is used for its analysis. WEKA allows the application of PCA and feature selection over the data, before the different classification algorithms are employed.

Table 4.11: All 122 Attributes in Our Data Set

1)**HAPPYSADLEVEL** {1,2,3,4,5,6,7,8},
2)**HAPPYRESULT** {1,2,3},
3)**HAPPYDESC** {1.1,1.2,1.3,1.4,2.1},
4)**SADDESC** {1.1,1.2,1.3,1.4,2.1},
5)HOW REAL,
6)STATE {stationary,moving},
7)**STATIONARYCOUNT** REAL,
8)**MOVINGCOUNT** REAL,
9)**UNIQUEAPS** REAL,
10)**UNIQUEOUICOUNT** REAL,
11)**LOCATIONBEENCOUNT** REAL,
12)DAYSONLINE_QTY REAL,
13)VIEWS_QTY REAL,
14)CONTRIBUTIONS REAL,
15)QUESTIONS REAL,
16)NOTES REAL,
17)ANSWERS REAL,
18)ASSIGNEDLECTURE REAL,
19)GRADELETTER {A,A-,B+,B,B-,E},
20)SLEEPINGHOUR REAL,
21)RATEAVRG REAL,
22)SOCIALAVRG REAL,

41)COHEN REAL,
42)COLLEGE-STREET REAL,
43)CUMMINGS REAL,
44)CURRIER REAL,
45)CUTTER-NORTH REAL,
46)DANA-LIBRARY REAL,
47)DARTMOUTH_HALL REAL,
48)DCCCC REAL,
49)DEWEY REAL,
50)EAST-WHEELOCK REAL,
51)EVERGREEN REAL,
52)EXTERNAL REAL,
53)EXTERNAL_25LEBANON REAL,
54)FAHEY-MCLANE REAL,
55)FAIRBANKS REAL,
56)FAIRCHILD REAL,
57)FAYERWEATHER REAL,
58)FELDBERG_LIBRARY REAL,
59)FRENCH REAL,
60)GILE REAL,
61)HALLGARTEN REAL,
62)HANOVERINN REAL,
63)HANOVERPSYCH REAL,
64)HILLCREST REAL,
65)HITCHCOCK REAL,
66)HOPKINS REAL,
67)ISR_WIRELESS REAL,
68)JUDGE REAL,

87)OCCUM REAL,
88)PARKHURST REAL,
89)PRESIDENTS_HOUSE REAL,
90)RAVEN-HOUSE REAL,
91)REED REAL,
92)REMOTE_OFFICES_HREAP REAL,
93)REMSEN REAL,
94)RICHARDSON REAL,
95)RIPLEY REAL,
96)ROBINSON REAL,
97)ROLLINS-CHAPEL REAL,
98)ROPEFERRY REAL,
99)SANBORN REAL,
100)SILSBY-ROCKY REAL,
101)SMITH REAL,
102)SOFTBALLFIELD REAL,
103)SPHINX REAL,
104)SPORT-VENUES REAL,
105)SPORT-VENUES-PRESS REAL,
106)STEELE REAL,
107)STREETER REAL,
108)SUDIKOFF REAL,
109)THAYER_SECURE REAL,
110)THORNTON REAL,
111)TLLC REAL,
112)TLLC-RAETHER REAL,
113)TOPLIFF REAL,

23)35CENTERRA REAL, 24)53_COMMONS REAL, 25)7-LEBANON REAL, 26)AQUINAS REAL, 27)BAKERBERRY REAL, 28)BATRLETT REAL, 29)BERRY_SPORTS_CENTER REAL, 30)BISSELL REAL, 31)BLUNT_ALUMNI_CENTER REAL, 32)BROWN_HALL REAL, 33)BUCHANAN REAL, 34)BURKE REAL, 35)BUTTERFIELD REAL, 36)BYRNEHALL REAL, 37)CARPENTERHALL REAL, 38)CARSON-TECH_SERVICES REAL, 39)CHANNING-COX REAL, 40)CHASEHALL REAL,	69)KELLOGG REAL, 70)KEMENY REAL, 71)LIBRARY-DEFAULT-SERVICES REAL, 72)LITTLE_HALL REAL, 73)LODGE REAL, 74)LORD REAL, 75)LSB REAL, 76)MACLEAN REAL, 77)MASSROW REAL, 78)MAXWELL REAL, 79)MCKENZIE REAL, 80)MCLAUGHLIN REAL, 81)MCNUTT REAL, 82)MOORE REAL, 83)MURDOUGH REAL, 84)NEWHAMP REAL, 85)NORTH-MAIN REAL, 86)NORTH-PARK REAL,	114)TUCK_HALL REAL, 115)VAC REAL, 116)VAIL REAL, 117)WEBSTERHALL REAL, 118)WENTWORTH REAL, 119)WHEELER REAL, 120)WHITTEMORE REAL, 121)WOODBURYHALL REAL, 122)WOODWARD REAL

The bold field attributes are result of feature engineering.

5. EVALUATION OF THE IMPACT OF MOBILITY ON MOOD

In this chapter, the effect of different parameters (such as classifier, feature selection) on the mood classification accuracy is iteratively analyzed.

The classification of StudentLife is done on mood (happy/sad) dataset using three different algorithm, which were explained in Section 4.2.4.

5.1 RESULTS WITHOUT ATTRIBUTE SELECTION

First, we investigate the accuracy of mood classification using different algorithms, when no attribute selection or PCA were applied. To observe the impact of mobility = on the mood, the classification methods are applied without and with the location/mobility attributes. The results of these experiments indicate that the accuracy of mood classification increases when mobility and location-related attributes are incorporated, compared to when they are not. The confusion matrix corresponding to the most accurate results without mobility data, without attribute selection is shown for Random Forest in Table 5.1.

Table 5.1: Detailed result of cross validation of Random Forest Classifier without attribute selection and with mobility data

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.832	0.465	0.717	0.832	0.770	0.387	0.729	0.761	1
	0.500	0.097	0.617	0.500	0.552	0.435	0.768	0.609	2
	0.209	0.109	0.290	0.209	0.243	0.114	0.583	0.249	3
Weighted Avrg.	0.643	0.315	0.618	0.643	0.626	0.351	0.713	0.634	

In table 12, we present the detailed breakdown of classification results over mood variables (happy: 1, sad: 2 and neither happy nor sad: 3). Also the f-measure result shows that class-3 variables had the lowest value compared to classes 1 and 2. The table shows the greatest precision on happy decision to classify as 1. The poorest precision is taken for neither happy nor sad classified variables. This may be due to neither happy nor sad variables being few in count compared to the rest of the dataset.

Table 5.2: Confusion Matrix classification of Random Forest without attribute selection and with mobility patterns

a	b	c	Classified As
119	10	14	a=1
21	29	8	b=2
26	8	9	c=3

When attribute selection or PCA are not applied, the most accurate results with location are obtained using the Random Forest algorithm. We present the corresponding detailed breakdown and confusion matrix in Table 5.1 and Table 5.2.

Table 5.3: Detailed Accuracy of happyresult classifier (1, 2, and 3) by using Random Forest Classifier without attribute selection without mobility patterns

	TP Rate	FP Rate	Precision	Recall	F-Meausure	MCC	ROC Area	PRC Area	Class
	0.846	0.644	0.651	0.846	0.736	0.234	0.644	0.701	1
	0.431	0.091	0.595	0.431	0.500	0.383	0.712	0.557	2
	0.023	0.075	0.063	0.023	0.034	-0.079	0.533	0.191	3
Weighted Avrg.	0.602	0.412	0.534	0.602	0.556	0.214	0.641	0.577	

Table 5.4: Confusion Matric classification of Random Forest classified with happyresult without attribute selection and without mobility patterns.

a	b	c	Classified As
121	11	11	a=1

29	25	4	b=2
36	6	1	c=3

When attribute selection or PCA are not applied, the most accurate results with location are obtained using the Random Forest algorithm. We present the corresponding detailed breakdown and confusion matrix in Table 5.3 and Table 5.4. The result also shows that without attribute selection, PCA and mobility patterns has higher prediction and weighed F scores in Table 5.5.

Table 5.5: Performance with/without Location information.

	j48	**Random Forest**	**SVM**	**MLP**
With Location	Accuracy 50.4098 % F-Measure 0.492 TPR 0.504 Precision 0.481	Accuracy 60.6557 % F-Measure 0.559 TPR 0.607 Precision 0.538	Accuracy 58.6066 % F-Measure - TPR 0.586 Precision -	Accuracy 53.2787 % F-Measure 0.533 TPR 0.533 Precision 0.534
Without Location	Accuracy 60.2459 % F-Measure 0.562 TPR 0.602 Precision 0.559	Accuracy 65.1639 % F-Measure 0.640 TPR 0.652 Precision 0.634	Accuracy 58.6066 % F-Measure - TPR 0.586 Precision -	Accuracy 60.2459 % F-Measure 0.586 TPR 0.602 Precision 0.578

5.2 ATTRIBUTE SELECTION

Next, we analyze the impact of Feature Selection on the classification accuracy of mood, with mobility data. The class is selected as happyresult, with possible values being 1, 2, 3 (1: Happy, 2: Sad, 3: Neither Happy nor Sad).

The first attribute selection method applied over the dataset was **cfsSubSetEval.** The algorithm reduced the 122 attributes to the following 7: HOW, VIEWS_QTY, NOTES, RATEAVRG, SOCIALAVRG, HOPKINS, BLUNT_ALUMNI_CENTER. The *HOW* column represents the mood, *VIEWS_QTY* represents Piazza usage and *RATE_AVRG, SOCIAL_AVRG* variables show sleep quality levels. The last two columns represent places in Dartmouth campus.

We observe that 2 of the selected subset of 7 attributes out of the total 122 attributes are location attributes; which is deemed as location data has an impact on the students' moods. The two most influential locations are the Blunt Alumni Center and Hopkins Center for the Arts. As we do not exactly know the size or profile of these two places, we cannot perform further interpretation on the mobility patterns to infer whether students visit or walk by these places, or why they do so.

The second feature selection method is **classifiersubsetEval**. We observe that it selects *HOW*, *UNIQUEAPS*, *DAYONLINE_QTY*, *CONTRIBUTIONS*, *SLEEPINGHOUR*, *FRENCH*, *HOPKINS*, *LORD* and *REED*. Again, 4 out of 9 (almost 50%) of the selected attributes are location attributes; which demonstrates to us an impact of location on the mood of the students.

Comparing with the cfsSubSetEval, this feature selection selected *UNIQUEAPS* which gives daily unique access point counts and *CONTRIBUTIONS* from Piazza usage, as well as including three more places *FRENCH*, *LORD* and *REED*. The one single column that is selected by both feature selection algorithms is *HOPKINS*, it is followed by *HOW* and *RATEAVRG* values. Table 5.6 contains descriptions of these selected features.

Table 5.6: The selected attributes' descriptions.

Attributes	**Descriptions**
HOW	Mood2 data set EMA question's answer "How do you feel now?". The answer is taken daily.
VIEWS_QTY	The student daily piazza application usage quantity.
NOTES	The count of how many notes are taken on the Piazza system.
RATEAVRG	The daily sleeping EMA question's answer. "How would rate your overall

	sleep last night?”
SOCIALAVRG	The daily sleeping EMA question’s answer. “How often did you have trouble staying awake yesterday while in class, eating meals or engaging in social activity?"
HOPKINS	The WiFi location place name is labeled.
BLUNT_ALUMNI_CENTER	The WiFi location place name is labeled.

Table 5.7: Attribute selection method results

Attribute Selection Method	**Selected Attributes**	**Method**
cfsSubSetEval	HOW, VIEWS_QTY, NOTES, RATEAVRG, SOCIALAVRG BLUNT_ALUMNI_CENTER, HOPKINS	Best first
classifiersubsetEval	2,6,9,11,17,56,63,71,88 : 9 HOW, UNIQUEAPS, DAYSONLINE_QTY, CONTRIBUTIONS, SLEEPINGHOUR, FRENCH, HOPKINS, LORD, REED	GreedyStepwise

After feature selection was complete, all classification algorithms were executed for the “HappyResult” class. The selected attributes shown in Table 5.8 have been used. J48, Random Forest and SVM algorithms have been executed sequentially on this data set. The Random Forest algorithm gave the highest accuracy of correlation at 63.1148%. Also we observed the result with different selected columns such as *HOW, VIEWS_QTY, NOTES, RATEAVRG, SOCIALAVRG, BLUNT_ALUMNI_CENTER, HOPKINS* and the accuracy result is approximately 60%.

As a result, on the research is observed that the selection of attributes with *HAPPYRESULT* classification increased the accuracy to 63.1148%. In addition, two of attributes selection the Random Forest is the most successful algorithm.

Table 5.8: After attribute selection is applied with data mining algorithms result chart

Algorithm	Attributes	Correctly Classified
J48	HAPPYRESULT, HOW, UNIQUEAPS, DAYSONLINE_QTY, CONTRIBUTIONS, SLEEPINGHOUR, FRENCH, HOPKINS, LORD, REED	61.0656 %
J48	HAPPYRESULT, HOW, VIEWS_QTY, NOTES, RATEAVRG, SOCIALAVRG, BLUNT_ALUMNI_CENTER, HOPKINS	61.0656 %
Random Forest Algorithm	HAPPYRESULT, HOW, UNIQUEAPS, DAYSONLINE_QTY CONTRIBUTIONS, SLEEPINGHOUR, FRENCH, HOPKINS, LORD REED	62.2951 %
Random Forest Algorithm	HAPPYRESULT, HOW, VIEWS_QTY, NOTES, RATEAVRG, SOCIALAVRG, BLUNT_ALUMNI_CENTER, HOPKINS	63.1148 %
SVM	HAPPYRESULT, HOW, VIEWS_QTY, NOTES, RATEAVRG, SOCIALAVRG, BLUNT_ALUMNI_CENTER, HOPKINS	61.0656 %
SVM	HAPPYRESULT, HOW, UNIQUEAPS, DAYSONLINE_QTY, CONTRIBUTIONS, SLEEPINGHOUR, FRENCH, HOPKINS, LORD, REED	60.6557 %

The most accurate, attribute selection with location confusion matrix's result is shown for MLP algorithm.

Table 5.9: Detailed Accuracy of happyresult classifier (1, 2, and 3) by using MLP Classifier with attribute selection with mobility patterns

	TP Rate	FP Rate	Precision	Recall	F-Meausure	MCC	ROC Area	PRC Area	Class
	0.916	0.535	0.708	0.916	0.799	0.439	0.708	0.719	1
	0.552	0.102	0.627	0.587	0.587	0.471	0.778	0.591	2
	0.140	0.010	0.750	0.235	0.235	0.277	0.550	0.266	3
Weighted Avrg.	0.693	0.339	0.696	0.649	0.649	0.418	0.697	0.609	

Table 5.10: Confusion Matrix classification of MLP classified with happyresult with attribute selection and with mobility patterns.

a	b	c	Classified As
131	11	1	a=1
25	32	1	b=2
29	8	6	c=3

When attribute selection applied, the most accurate results with location are obtained using the MLP algorithm. We present the corresponding detailed breakdown and confusion matrix in Table 5.9 and Table 5.10. The result also shows that with attribute selection and mobility patterns has higher prediction and weighed F scores. Also selected attributes are highly mobility patterns which shows the relation between mobility and mood attributes.

The most accurate, attribute selection without location confusion matrix's result is shown for Random Forest algorithm.

Table 5.11: Detailed Accuracy of happyresult classifier (1, 2, and 3) by using Random Forest Classifier with attribute selection without mobility patterns

	TP Rate	FP Rate	Precision	Recall	F-Meausure	MCC	ROC Area	PRC Area	Class
	0.797	0.475	0.704	0.797	0.748	0.336	0.765	0.794	1
	0.500	0.118	0.569	0.500	0.532	0.400	0.771	0.616	2
	0.140	0.124	0.194	0.140	0.162	0.017	0.676	0.266	3
Weighted Avrg.	0.693	0.339	0.696	0.649	0.649	0.295	0.751	0.658	

Table 5.12: Confusion Matric classification of Random Forest classified with happyresult with attribute selection and without mobility patterns.

a	**b**	**c**	**Classified As**
114	12	17	a=1

21	29	8	b=2
27	10	6	c=3

The most accurate, attribute selection without location confusion matrix's result is shown for Random Forest algorithm result and the most accurate attribute selection algorithm with location result compared, it is shown that F – measure result and precision for classification over happy had better results if we compared with sad or neither happy nor sad result. Also attribute selection with location result overall results shows that had better results on MLP at Table 5.11 and 5.12.

5.3 IMPACT OF APPLYING PCA

The PCA is applied at pre-process for seeing the affection of the algorithm. The class is chosen as HAPPYRESULT. The data is normalized firstly then PCA is applied.

After PCA, the classification methods are applied orderly. The results is shown at Table 5.13.

Table 5.13: After PCA applied the classification result chart

Algorithm	Attributes	Correctly Classified
J48	*all merged data	52.0492 %
Random Forest – 100 Iterations	*all merged data	59.8361 %
Random Forest – 500 Iterations	*all merged data	61.0656 %
SVM – Radial Basis Kernel	*all merged data	61.4754 %
SVM – Sigmoid Kernel	*all merged data	62.2951 %

In sum, with feature selection or without the relationship between mood and mobility is detected. Also created one more table for clarifying location effect on the students' mood. It is shown at Table 5.14.

Table 5.14: The result of without mood attributes

Algorithm	Attributes	Correctly Classified
J48	HAPPYRESULT DAYSONLINE_QTY VIEWS_QTY CONTRIBUTIONS QUESTIONS NOTES ANSWERS ASSIGNEDLECTURE GRADELETTER SLEEPINGHOUR RATEAVRG	59.4262 %
Random Forest	HAPPYRESULT DAYSONLINE_QTY VIEWS_QTY CONTRIBUTIONS QUESTIONS NOTES ANSWERS ASSIGNEDLECTURE GRADELETTER SLEEPINGHOUR RATEAVRG	64.3443 %
SVM	HAPPYRESULT DAYSONLINE_QTY VIEWS_QTY CONTRIBUTIONS QUESTIONS NOTES ANSWERS ASSIGNEDLECTURE GRADELETTER SLEEPINGHOUR RATEAVRG	63.5246 %

The PCA application with and without location confusion matrix's result is shown for the most accurate algorithm SVM for both cases.

PCA with location on SVM:

Table 5.15: Detailed Accuracy of happyresult classifier (1, 2, and 3) applied PCA by using SVM Classifier with mobility patterns

	TP Rate	FP Rate	Precision	Recall	F-Meausure	MCC	ROC Area	PRC Area	Class
	0.937	0.812	0.620	0.937	0.747	0.193	0.563	0.618	1
	0.259	0.048	0.625	0.259	0.366	0.301	0.605	0.338	2
	0.000	0.020	0.000	0.000	0.000	-0.060	0.490	0.176	3

Weighted Avrg.	0.611	0.491	0.512	0.611	0.524	0.174	0.560	0.474	

Table 5.16: Confusion Matric classification of SVM classified with happyresult with PCA and with mobility patterns

a	b	c	Classified As
134	6	3	a=1
42	15	1	b=2
40	3	0	c=3

PCA without location on SVM:

Table 5.17: Detailed Accuracy of happyresult classifier (1, 2, and 3) applied PCA by using SVM Classifier without mobility patterns

	TP Rate	FP Rate	Precision	Recall	F-Meausure	MCC	ROC Area	PRC Area	Class
	0.944	0.703	0.655	0.944	0.774	0.327	0.621	0.651	1
	0.362	0.032	0.778	0.362	0.494	0.448	0.665	0.433	2
	0.047	0.045	0.182	0.047	0.074	0.003	0.501	0.176	3
Weighted Avrg.	0.648	0.428	0.601	0.648	0.584	0.299	0.610	0.516	

Table 5.18: Confusion Matric classification of SVM classified with happyresult with PCA and with mobility patterns

a	b	c	Classified As
135	2	6	a=1
34	21	3	b=2
37	4	2	c=3

In the confusion matrix results shows that both results are really closer, different than accuracies. In conclusion, we think that PCA algorithm is not matching our dataset for applying for. We could not get any expected results with location or without location. We also observed that mobility with happy has relation with each other. All the confusion matrix chart of results are shown Table 5.15, 5.16, 5.17, and 5.18.

5.4 CHECKING THE RESULTS FOR OVERFITTING

In this thesis, the WEKA default parameters are highly used. All the algorithms' results are taken by cross-validation with k-fold (10 folds). MLP algorithm with attribute selection is repeated for 20, 30, 40 folds. The accuracy decreases to 64.7541 % when we changed the fold to 20. The 30 fold gives better result than 20 fold, for which the accuracy is 65.9836 %. The 40 fold is also give same result with 20 fold. However any of them is not successful as much as 10 fold.

The percentage splits are also applied to the MLP with attribute selection and the result of percentage split is viewed for 60%, 70% and 80%. The 80% result is 73.4694 %, 70% result is 72.6027 % and 60% result is 65.3061 %. In this case we assume that without cross validation the result of accuracies are increased when we used the percentage splitting with the training set. That is why we can say that there is an over fitting.

We applied MLP algorithm with the training set to the full dataset. The training time default 500ms is applied then 800ms and 1000ms are applied for seeing result effects. The default parameter 500 the accuracy is 73.7705 %. The 800’s accuracy is 75.4098 % and the final training time parameter 1000 result is 75.4098 %. After attribute selection, there is not much attributes exist for creating dataset, that is why the other higher result algorithm of Random Forest with location variables without attribute selection is worked over a training set. The result is really interesting that when applied to the training set, the accuracy is react to the 97.541 % which is really high result if it will be compared with cross-validation results. The iterations variables are also changed however accuracy is not increased.

Summary of Findings:

Table 5.19: The conclusion results of all algorithms

Attribute Selection	j48	Random Forest	SVM	MLP
No Attribute Selection, *Without* Location	Accuracy: 60.2459 % F-Measure: 0.562 TPR: 0.602 Precision: 0.559	Accuracy: 65.1639 % F-Measure: 0.640 TPR: 0.652 Precision: 0.634	Accuracy: 58.6066 % F-Measure: ? TPR: 0.586 Precision: ?	Accuracy: 60.2459 % F-Measure: 0.586 TPR: 0.602 Precision: 0.578
No Attribute Selection, *With* Location	Accuracy: 50.4098 % F-Measure: 0.492 TPR: 0.504 Precision: 0.481	Accuracy: 60.6557 % F-Measure: 0.559 TPR: 0.607 Precision: 0.538	Accuracy: 58.6066 % F-Measure: ? TPR: 0.586 Precision: ?	Accuracy: 53.2787 % F-Measure: 0.533 TPR: 0.533 Precision: 0.534
Attribute Selection Applied, *Without* Location	Accuracy: 61.8852 % F-Measure: 0.576 TPR: 0.619 Precision: 0.580	Accuracy: 62.7049 % F-Measure: 0.612 TPR: 0.627 Precision: 0.603	Accuracy: 58.6066 % F-Measure: ? TPR: 0.586 Precision: ?	Accuracy: 57.377 % F-Measure: 0.554 TPR: 0.574 Precision: 0.543
Attribute Selection Applied, *With* Location	Accuracy: 61.4754 % F-Measure: 0.594 TPR: 0.615 Precision: 0.585	Accuracy: 63.5246 % F-Measure: 0.649 TPR: 0.635 Precision: 0.599	Accuracy: 58.6066 % F-Measure: ? TPR: 0.586 Precision: ?	Accuracy: 67.2131 % F-Measure: 0.649 TPR: 0.672 Precision: 0.660
PCA Applied, *Without* Location	Accuracy: 60.2459 % F-Measure: 0.562 TPR: 0.602 Precision: 0.559	Accuracy: 61.4754 % F-Measure: 0.595 TPR: 0.615 Precision: 0.586	Accuracy: 64.3443 % F-Measure: 0.593 TPR: 0.643 Precision: 0.615	Accuracy: 61.0656 % F-Measure: 0.590 TPR: 0.611 Precision: 0.581

PCA Applied, *With* Location	Accuracy: 45.4918 % F-Measure: 0.454 TPR: 0.455 Precision: 0.453	Accuracy: 59.8361 % F-Measure: 0.541 TPR: 0.598 Precision: 0.523	Accuracy: 61.0656 % F-Measure: 0.516 TPR: 0.611 Precision: 0.520	Accuracy: 48.3607% F-Measure: 0.483 TPR: 0.484 Precision: 0.483

6. DISCUSSION AND CONCLUSION

This thesis focused on analyzing the impact of students' mobility patterns on their mood using classification techniques. For this purpose, we utilized the StudentLife dataset from the literature.

The parameters of StudentLife dataset that include Piazza usage, sleeping and Wi-Fi are used for analyzing the effect on the mood. First of all, the mood data result may have different results between perceived and reality of happiness level. Therefore for each student the mood review should be personalized. The real time physiological sensors can work effectively in these cases. It can be good supporter for following research which data does not exist for the current thesis.

Over the result of the analysis, clearly obtained that happiness is related to Piazza usage, sleeping and mobility.

The impact of classifiers, PCA and attribute selection have been iteratively studied. To evaluate the impact of mobility patterns on the mood classification, we have calculated the results with and without the location attributes.

First, attribute selection is applied with location and without location data set. The results shows that attribute selection with location is giving better results than without location results. The highest accuracy gets on MLP classifier, at 67.2131 % accuracy on attribute selection with location. The MLP is giving really good results on many kinds of datasets. MLPs are suitable for classification prediction problems where inputs are assigned a class or label that in thesis the labeled attribute is HAPPYRESULT. They are very flexible and can be used generally to learn a mapping from inputs to outputs. After analyzing all the chart, the highest weighed f-measure value is shared with two algorithms MLP and Random Forest on 0,649.

Then, attribute selection is applied with PCA with and without location data set. However, the result is different than attribute selection with PCA for MLP classifier. It gives lowest accuracy, 48.3607% for PCA with location. The SVM classifier gets the highest accuracy, at 64.3443 % on PCA when eliminated mobility attributes over the dataset.

Finally, attribute selection is not applied on the dataset, to observe its impact on mood classification accuracy with and without the location information. The general accuracy over the classifier were really closer. But if the result are compared J48 algorithm gets the lowest accuracy, 50.4098% for with location data. At the same time without location data the lowest accuracy, 58.6066 % is taken by SVM classifier. The highest accuracy of 65.1639 % is taken by Random Forest algorithm when location parameters do not selected.

In conclusion, all above of the classifiers, the lowest accuracies are generally taken on J48 classifier. The most successful classifier results are taken on Random Forest classifier. However the highest accuracy is taken on MLP classifier on attribute selection with location is taken.

The PCA application on data is not making a huge difference in accuracy. However attribute selection is affecting results positively.

The results obtained in this study can be helpful in investigating the following student life overviews and more featured data collections. On the future work can view more specific mood data with more students on different patterns for analyzing the connection with student’s mood.

REFERENCES

Books

Rappaport, T. S. (2002). *Wireless communications: Principles and practice*. Upper Saddle River, N.J: Prentice Hall PTR.

Periodicals

Gjoreski, M. & Gjoreski, H. & Lutrek, M. & Gams, M. (2015) Automatic Detection of Perceived Stress in Campus Students Using Smartphones, 2015. *International Conference on Intelligent Environments*, pp. 132-135.

A. Ghandeharioun et al. (2017) *Objective assessment of depressive symptoms with machine learning and wearable sensors data*. In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pages 325–332, 2017.

R. Wang et al. (2013) Studentlife: *Assessing mental health, academic performance and behavioral trends of college students using smartphones*. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp ’14, pages 3–14, 2014.

Other Publications

M. S. H. Aunget al. (2016) *Leveraging multi-modal sensing for mobile health: A case review in chronic pain. IEEE Journal of Selected Topics in Signal Processing*, 10(5):962–974, 2016.

Saeb, S. & Lattie, E.G.. & Schueller, S.M. & Kording, K.P. & Mohr, D.C. (2016) *The relationship between mobile phone location sensor data and depressive symptom severity, 2016*. *PeerJ*, 4.

D. Ben-Zeev et al. (2017) Crosscheck: *Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. Psychiatric rehabilitation journal*, 40(3):266–275, 2017.

G.M. Harari et al. (2017) *Patterns of behavior change in students over an academic term. Comput. Hum. Behav*., 67(C):129–138, February 2017.

W. Wang et al. (2018) Sensing behavioral change over time: *Using within-person variability features from mobile sensing to predict personality traits. In Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, volume 2, pages 141:1–141:21, 2018.

S.M. Schueller K.P. Kording D.C. Mohr S. Saeb, E.G. Lattie. (2016) *The relationship between mobile phone location sensor data and depressive symptom severity*. PeerJ., 4, 2016.

K.B.Lipkowitz T.R.Cundari (2007) *Applications of Support Vector Machines in Chemistry, Rev. Comput. Chem*. 2007, 23, 291-400

Jadi, Z. & Steven M. & Smith,L. I. (2018) *A step by step explanation of Principal Component Analysis*, 2018

Wang,R. & Chen,F. & Chen,Z. & Li,T. & Harari, G. & Tignor,S. & Zhou,X. & Ben-Zeev,D. & Campbell,T.A. (2018) *Assessing Mental Health, Academic Performance and Behavioral Trends of College Students using Smartphones*,2014.

Yiu,T. (2018) *Understanding Random Forest*

M. A. Hall. *Correlation-based Feature Subset Selection for Machine Learning. Hamilton*, New Zealand,1998.

Lin H., *A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods*,2003

Chen Z.,Lin M.,Chen F.,D. Lane N.,Cardone G et al. (2013), *Unobtrusive Sleep Monitoring using Smartphones,*May 2013

Boukhechba M., Fua K. et al.,(2017) *Monitoring Social Anxiety from Mobility and Communication Patterns,*(September 2017)

Wa L., Liu Y.,D. Lane N.,Zhong L.,2013, *MoodScope: Building a Mood Sensor from Smartphone Usage Patterns Robert* ,(2013)

David C. Mohr, Mi Zhang, and Stephen M. Schueller. 2017., *Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. Annual Review of Clinical Psychology,* 13, 1 (May 2017)

Lindsay I. Smith (2002), *A tutorial on Principal Component Analysis,* Pages 12-17, February-2002