

# Wi-Fi Feature Engineering for Detection of Campus Social Dynamics and Academic Performance Prediction

Devashru Patel

Faculty Member #1:

Printed Name: Gregory Abowd

---

Signature: 

---

Faculty Member #2:

Printed Name: Thomas Ploetz

---

Signature: 

---

# 1 Introduction

University students are some of the most susceptible individuals to stress-related mental health issues. This stress could stem from a variety of factors including social relationships, familial issues, and also course workload. Universities have noted these issues and the most common method for identifying such students is through the use of mental health surveys periodically throughout the academic term. The challenge that arises with this approach is that individuals may often put false answers on such surveys, and they are also quite difficult to conduct on a large scale due to the difficulty of enforcing all students to take them. Recently however, the StudentLife study was conducted in which students were requested to provide their smartphone sensor data as a means of assessing mental health. The results found that there is “significant correlation between objective sensor data from smartphones and outcomes from validated mental health ‘gold standard’ surveys” (Wang et al. 2014). Sensor data can include location data or even microphone input, which people are often reluctant to provide access to as such data is considered quite invasive. Our motivation as a team is to exclusively use Wi-Fi data to preemptively identify when a student is experiencing some form of negative stress and be able to intervene to help alleviate any problems.

Researchers have begun using smartphones as a new medium for research into student interaction. One team utilized accelerometer and microphone data from student’s phones to determine levels of activity and sociability. Data was collected over a 10 week period and over that time “stability estimates for weekly activity and sociability were moderate to high” (Harari et al. 2017). This implies that the patterns of movement and peer interaction that a student exhibits early in the semester is highly likely to transfer to the end of the semester as well. This

confirms that such a method is a reliable means to predict individual behavior as the manner in which an individual acts will persist throughout an academic term.

Most studies such as Harari's require downloading an application and allowing many permissions such as access to location data or microphone input. However, many students are often reluctant to provide such access due to the implication of an invasion of privacy. An alternative method that is less invasive, but that may not provide as much information, is the use of Wi-Fi data. Often researchers can use data from the central network at an institution because it is easily accessible with only simple consent from the student being required to proceed. One study chose to employ "density-based clustering of WLAN traces" to "automatically detect regular learning activities, and discover social groups among the students, who attend these activities" (Eldaw, Levene, and Roussos 2017). The locations of classes that students were known to attend at certain times were used to validate the accuracy of their estimations.

Our project CampusLife and its methods have been greatly influenced by the ones provided in the Eldaw and StudentLife papers. The assumption is that a student who is experiencing better social dynamics with their peers, physical exercise, and overall has better mental health tends to perform better academically (Eisenberg, Golberstein, and Hunt 2009). To accomplish our task, a class of students was asked to provide access to their Eduroam logs and those who obliged became part of this study. Each student must sign in with their own credentials so finding their logs in the system is quite simple. What Eduroam logs provide is when and which internet access point each student accessed as they were going about their day. As the locations of access points are known, this allows us as researchers to gather a generally

good idea of where students are located throughout the day and how often they are with other students in the class. The data will then be fed into machine learning models to predict what a student's academic performance will be at the end of the term. We hope to use these models to then intervene as early as possible by either recommending resources or counseling to help alleviate mental health issues and improve academic performance. Smartphone technology has a big step forward for mental health and utilized in various other studies. Our study seeks to broaden the horizons of the use of such a technology in helping student mental health.

## **2 Literature Review**

Helping improve the academic outcomes of students has been the main objective of countless educational psychology studies. While it is quite evident that resources inside the classroom environment can have a huge impact upon a student's performance, the influence of factors outside the classroom should not be neglected. The "occurrence of depression and anxiety is associated with a significant additional drop in GPA" (Eisenberg, Golberstein, and Hunt 2009). Positive and negative emotions alone can explain much of an individual's current emotional outlook relating to issues such as depression or anxiety. This is why the Positive and Negative Affect Schedule (PANAS) questionnaire was created decades ago and is still often used in diagnosing mental health conditions (Watson, Clark, and Tellegen 1988). However, it would be difficult to provide questionnaires to all students at a university without the use of some online system. In addition, respondents tend to lie in self-reported surveys and overreport their own success or habits (Brenner and DeLamater 2016). Another way to understand mental health that may not be as direct, but still prove valuable is looking at student and peer behavior.

A novel method for studying student behavior is the use of Wi-Fi data to understand the movement of groups of people throughout a dense area and such techniques are referred to as density-based. A team headed by Muawya Eldaw developed and over time has made vast improvements to a density-based clustering algorithm named Social-DBSCAN . The algorithm extracts data from the wireless local area network (WLAN) and is able to automatically detect regular learning activities and from this infer attendance levels over the academic term with high accuracy (Eldaw, Levene, and Roussos 2017). It essentially finds when students are present in the same location at the same time to detect that a class session is occurring and see who is attending and absent based on such data. Given its success in predicting classroom activities, Social-DBSCAN could then be expanded for use in crowd detection during sports games or other non-academic events. The CampusLife study intends to expand similar algorithmic methods to the campus level and apply it in both scenarios inside of and outside of class.

All the density-based metrics provide valuable insight into the movement of students around a university campus, but they only provide information on groups as a whole. More individual data is required to truly help struggling students. Perhaps the most common approach that researchers have taken towards resolving this issue and gaining more information about the individual is by employing smartphone sensor data. One group created an application that participants would install on their devices and transmit location information would be relayed back for use in models (Dong, Lepri, and Pentland 2011). This paper cemented the idea that sensor data was capable of detecting social relationships amongst people and how it involved the behaviors of an individual as well. The StudentLife initiative at Dartmouth University later took this methodology and applied it on an even larger scale: accelerometer data that tracked

movement was used to perceive physical activity while the microphone was used to detect sociability (Wang et al. 2014). The results indicated that indeed the use of such sensory data was just as good as predicting academic performance as the data gathered through administering mental health surveys. Given that these surveys were the gold standard prior to the use of such technology, smartphone sensor analysis seems quite promising. The StudentLife team also found that personality characteristics such as stress of individuals in comparison to their peers remained relatively stable. This means that if a student was typically more stressed than their peers in the beginning of the semester, the same would hold true during midterm evaluations even though the stress level of the entire campus has risen as a whole (Harari et al. 2017). Using this knowledge, we know that students that experience a greater amount of anxiety earlier in the academic term are more at risk and we can target such individuals.

There are certain advantages to using sensory data that are undeniable. For example, accelerometer data could be utilized to indicate levels of activity such as jogging. However, smartphone sensor access is quite invasive and many people are likely to deny access to these permissions on their devices. A possible solution to this dilemma is by applying the aforementioned Wi-Fi analysis methods, but on the individual level. One group of researchers used WLAN traces from different access points to visualize movement throughout a hospital . As the access points are stationary and the locations known, a good estimate of the general location of anyone connected to a point can be made (Prentow et al. 2015).

This aforementioned WLAN trace technique is the methodology we have chosen to employ in our CampusLife project because it is the least invasive and constantly updates itself as

an individual connects to different access points as they move throughout the day. As we want to eventually expand this to a campus-wide system to monitor student mental health, remaining minimally invasive is vital to ensuring that students readily accept the system. However, it refines upon and adds much more to the examination of WLAN traces. A key concept useful for analyzing the relationships amongst people is the idea of social linkedness. Simply by seeing how often two individuals spend at the same Wi-Fi access point, we have a fair indication of how much social interaction they engage in (Jiang et al. 2015). We plan to incorporate this idea of social linkedness to see how often students interact with specific peers, but plan to make it even more robust by weighting times that individuals spend outside of class together more heavily as often this is more indicative of friendship. For our study, the time that participants spend in the course will be used primarily as a ground truth to ensure that our location predictions based on the WLAN traces are correct. We will see how students interact with their friends and peers on a daily basis and use this data to predict their final grade in the course.

### **3 Methodology**

The driving force behind our research is examining social interaction amongst university students. To accurately judge such a phenomenon is no easy feat, so the assumption was that the main indicator of social interaction amongst students can be quantified through collocations. Collocations in the context of our study are essentially students being present in the same location at the same time, which is information that can be gathered from the Wi-Fi data. The task of predicting academic performance then simplifies itself to identifying these collocations and seeing how well they correlate to academic success. The process that was followed to

accomplish this can be partitioned into three distinct sections: data collection, data manipulation, and prediction.

### **3.1 Gathering Data**

Legitimate network logs must first be obtained from a college campus to depict the behavior of real students. The logs must include info such as what Wi-Fi access point was accessed and when by any of a student's devices. The target group for our research was undergraduate students from a project-based course at the Georgia Institute of Technology. The required data was requested from these individuals and they all signed a consent form. The institute's IT department then provided anonymized network data for all of the student's various devices in which each student was represented by a code. The data spans across approximately 14 weeks or the duration of the entire semester. In addition, the class from which all students were recruited was cooperative, meaning that each student was given a score by the other students in their respective group correlating to how well they performed as teammates. This data was also gathered to see how effective our machine learning model was in predicting performance compared to more common methods such as peer evaluation.

### **3.1 Organizing and Analyzing the Information**

Collocations amongst students should only be acknowledged if the individuals are stationary and present at the same access point for an extended period of time. This avoids classifying situations such as when two students pass each other in a hallway as a collocation. Determining when a student is mobile or stationary then presents itself as being quite important.

#### **(i) Student in Motion**

Significant movement is always characterized by a change in the access point that a device connects to, but with the stipulation that it is done in a short period of time. The logs of each

device were examined individually and intervals between two consecutive log entries at different access points were measured. The 90th quantile value of this was used to determine movement. Any consecutive entries at different points that occurred within the calculated time were classified as in-motion segments.

**(i) Student is Stationary**

Any span of time that was not classified as a moving segment corresponds to a stationary segment.

**(i) Student Disconnected**

Many students leave the campus during various times of the day for reasons such as meals or returning to their off-campus residences. In terms of log entries this would result in one entry followed by the next entry being significantly later and possibly at a different access point as well. The only time at which we have the ground truth for student location is when they are attending the course from which they have been recruited. The longest interval between a device accessing a Wi-Fi point was then determined. As all devices in the class never exceeded this time before communicating with their tethered access point again, it is assumed that any segments longer than this indicate disconnections. Any point at which two log entries inside a stationary segment exceed the calculated maximum interval is labeled as a disconnection segment.

**(i) Students Collocated**

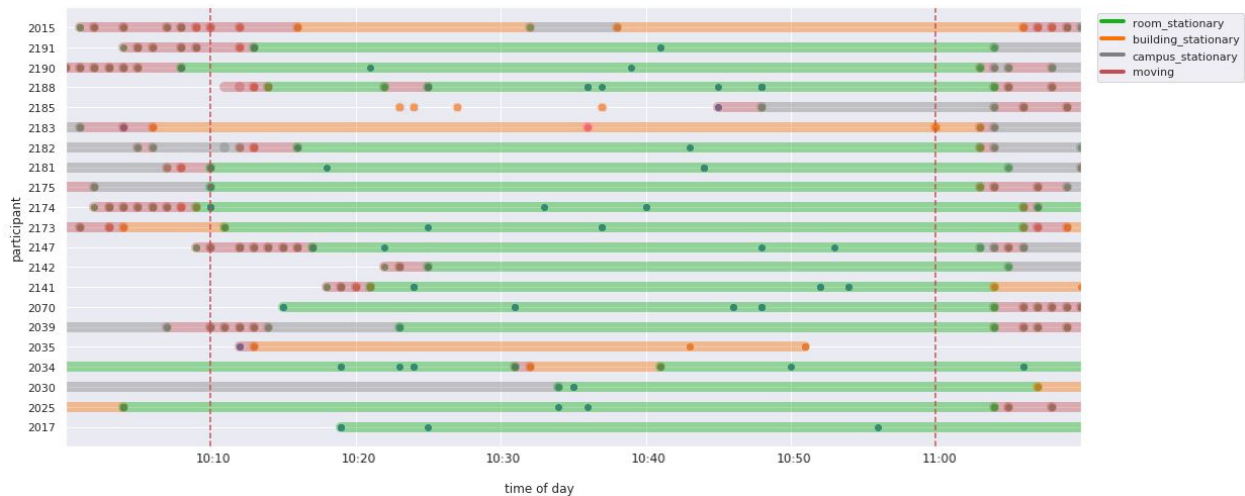
Overlapping stationary segments are seemingly the indicators for collocations. While this might be true, disturbances in Wi-Fi connection or simple restroom breaks should also be accounted for. The median time was taken for when overlapping stationary segments had such a gap. If the gap was smaller than the median, the collocation still exists despite the gap and is not interrupted.

### 3.3 Applying Machine Learning to Make Predictions

Besides collocations, various other factors that could present some sort of influence upon academic performance were also gathered (e.g., course schedule, building type). Features were extracted on a weekly level for each student as having them on a daily basis tends to cause too much variation and make the model pick up false trends. The features are split according to whether they pertain to individual or group activity. These feature sets are then fed into various machine learning models and used to make predictions for the final course grade that each student received. The predictions are then compared with the actual values to evaluate model performance. There will also exist a model trained upon just the peer evaluation scores to compare our models to the methods that were previously employed.

## 4 Results

The raw data had to first be analyzed to arrive at the stationary and moving segments. A visualization of the segments for each student in a class can be seen below:



The graphic shows a period of stationary segments surrounded by a period of moving segments. The results are promising as this is exactly what we would expect from students entering, attending, and then exiting a class.

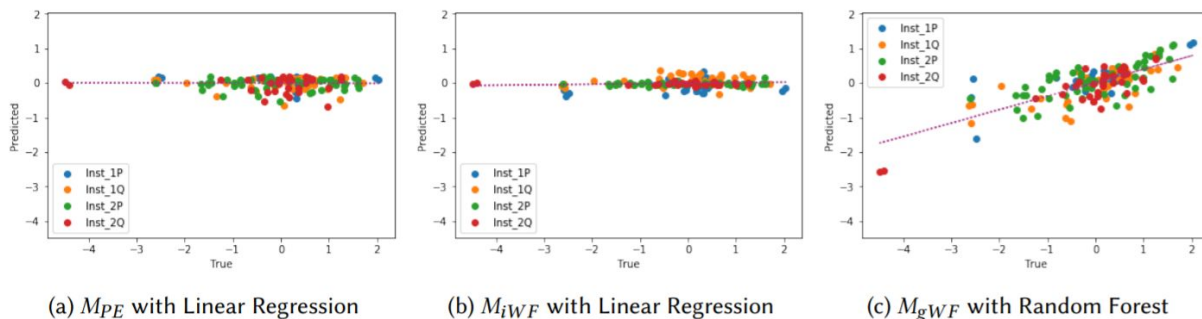
To achieve the best performance, multiple machine learning models were run such as linear regression and random forest. Whichever regressor led to the greatest reduction in RMSE from the baseline predictor was the model used for its respective training data. In this case, the baseline just involves outputting the mean of the target variable despite the input. All the information is presented in the following table:

<b>Training Data</b>	<b>Estimator</b>	<b>RMSE</b>	<b>Correlation</b>
-	Target Mean	1.07	-
Peer Evaluation	Linear Regression	1.08	-
Individual Features	Linear Regression	1.06	0.12
Group Features	Random Forest	0.74	0.75

The training data column for the first model remains empty as that was the baseline and was not actually trained, but rather just outputted the mean as detailed earlier. As is evident through the table, the best estimator was linear regression for the models trained upon the peer evaluation and individual feature datasets. For the group features, it was random forest that produced the most accurate predictions. Both the baseline and the model trained on the peer evaluation data had negligible correlation and therefore this statistic was excluded for the two models. The model trained on individual features was only marginally better with a slightly lower RMSE value and 0.12 correlation. The largest improvement came however, from the use of group features.

## 5 Discussion

The results are strongly in favor of group features such as time spent meeting with group members being the best predictor of academic performance. These findings are quite expected given that the course was project-based and would require coordination amongst peers to achieve success. The accuracy that the model utilizing group features produced far outclassed the use of both individual features and peer evaluation data as evidenced in the graphs below.



Of the three, peer evaluation data actually found itself to produce the most error, comparable to that of just predicting the mean in fact. This is in contrast to some previous findings, but in agreement with others. The results are not too perplexing as there is likely some social desirability bias inherent in all manual reporting of performance.

## 6 Conclusion

The purpose of our research was to predict the academic performance of students in a course early into the schedule of that course. Midterm evaluations are often proposed too late to be of any real use as the students that are struggling have fallen so far behind that they will experience immense difficulty in catching up. According to the results of our study, the result gained from such evaluations may not even be greatly indicative of academic performance as

they are prone to certain biases. So the concept we proposed was one of understanding social interaction and how it affects performance in school. Our hypothesis was that collocations between students was a suitable way to understand how often they interacted with each other as a group. After seeing collocation data vastly outperform individual Wi-Fi data and also the previous standard of peer evaluations, it is not farfetched to say that this is indeed the case. The use of Wi-Fi logs can be used to help identify those individuals on campus that are struggling or likely to struggle in their classes and be used as a means of early intervention.

## Citations

- Brenner, Philip S, and John DeLamater. "Lies, Damned Lies, and Survey Self-Reports? Identity as a Cause of Measurement Bias." *Social psychology quarterly* vol. 79,4 (2016)
- Dong, Wen, Bruno Lepri, and Alex (Sandy) Pentland. "Modeling the Co-Evolution of Behaviors and Social Relationships Using Mobile Phone Data." *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*, 2011.
- Eisenberg, Daniel, Ezra Golberstein, and Justin B. Hunt. 2009. "Mental Health and Academic Success in College." *B.E. Journal of Economic Analysis and Policy*, 9(1): Article 40.
- Eldaw, Muawya Habib Sarnoub, Mark Levene, and George Roussos. "Presence Analytics: Detecting Classroom-based Social Patterns using WLAN Traces." *Intelligent Systems Conference 2017*, 7-8 September 2017.
- Harari, Gabriella M., Samuel D. Gosling, Rui Wang, Fanglin Chen, Zhenyu Chen, and Andrew T. Campbell. "Patterns of Behavior Change in Students over an Academic Term: A Preliminary Study of Activity and Sociability Behaviors Using Smartphone Sensing Methods." *Computers in Human Behavior* 67 (2017): 129–38.
- Jiang, Shan, Xinning Zhu, Junfei Huang, and Guochu Shou. "Mining Social Groups in Campus Based on Wireless Detection." *2015 IEEE International Conference on Smart City/SocialCom/SustainCom (SmartCity)*, 2015.
- Kjaergaard, Mikkel Baun, Martin Wirz, Daniel Roggen, and Gerhard Troster. "Mobile Sensing of Pedestrian Flocks in Indoor Environments Using WiFi Signals." *2012 IEEE International Conference on Pervasive Computing and Communications*, 2012.
- Prentow, Thor S., Antonio J. Ruiz-Ruiz, Henrik Blunck, Allan Stisen, and Mikkel B. Kjærgaard. "Spatio-Temporal Facility Utilization Analysis from Exhaustive WiFi Monitoring." *Pervasive and Mobile Computing* 16 (2015): 305–16.
- Wang, Rui, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. "StudentLife." *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp 14 Adjunct*, 2014.

Watson, David, Lee Anna Clark, and Auke Tellegen. "Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales." *Journal of Personality and Social Psychology* 54, no. 6 (1988): 1063–70