# Predicting Students' Class Attendance

Maud Vissers
Master Thesis Data Science Business and Governance

THESIS SUBMITTED IN PARTIAL FULFILMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE OF DATA SCIENCE
AT THE FACULTY OF HUMANITIES
OF TILBURG UNIVERSITY

Thesis committee:
W. Huijbers
A.T. Hendrickson

Tilburg University
School of Humanities
Tilburg, The Netherlands
October 2018

# Preface

Writing this thesis completes my education at the Tilburg University, resulting in a Master of Science degree in Data Science. The completion of this thesis would not have been possible without the support and encouragement of the following people. First, I wish to express my gratitude to my supervisors Willem Huijbers and Drew Hendrickson for their meetings, guidance, and patience throughout the process of me writing this thesis. My sincere thanks goes to the researchers of the StudentLife study for tracking students for a period of ten weeks, for analyzing the information, and for authorizing public access to their dataset. I would like my fellow master students for the stimulating discussions. Finally, I would like to express my love and appreciation for all support by my family, friends, and in particular by Teun.

Maud Vissers, October 2018

# Abstract

This study is motivated by the lack of research on predicting class attendance in the field of pedagogy. Frequently, previous research conducted in the field of pedagogy concerned predicting students' performance using class attendance as predictor, where little research focused on predicting class attendance. The aims to investigate the probability of predicting class attendance for students' personal development, for professors' preparation and intervention, and to optimize universities' educational program are explained. The dataset obtained by the Studentlife study is used to conduct four experiments for predicting class attendance. The problem statement addressed in this thesis reads as follows: Can class attendance be predicted based on sensor data and education data? The techniques considered are Logistic Regression, Random Forest and Naïve Bayes. The main findings are that class attendance can be predicted based on sensor data and education data. The performance evaluated by the accuracy score depends on the algorithms and the predictors used. In this thesis, the best performing algorithm for predicting class attendance is the Random Forest algorithm containing GPS Location data, WiFi Location data, and Class Information data. Implications of the thesis contained having few participants, and the lack of homogeneous data. It would be interesting to test the proposed experiments on homogeneous data, and to use the proposed experiments for predicting class attendance based on GPS Location data obtained 24, 48, or 72 hours prior to the start of the class.

# Contents

**Preface** **1**

**Abstract** **2**

**Contents** **3**

**1. Introduction** **4**
1.1 Background 4
1.2 Problem Statement 6
1.3 Outline of the Thesis 7

**2. Related Work** **10**
2.1 Research Issue 10
2.2 Related Work 11
2.3 The Classifiers Used 14

**3. 3 Method** **18**
3.1 Dataset Information and Methodology 18
3.2 Data Cleaning 21
3.3 Feature Selection 22
3.4 Data Preparation and Partitioning 24
3.5 Experiments 24
3.6 Evaluation Method 28
3.7 Software 28

**4. 4 Experiments and Results** **29**
4.1 Results Experiment 1 30
4.2 Results Experiment 2 32
4.3 Results Experiment 3 34
4.4 Results Experiment 4 39

**5. General Discussion and Conclusions** **42**
5.1 Answers to the Research Questions 42
5.2 Directions for Further Research 47

**References** **49**

**Appendix** **52**

# Section 1: Introduction

This section explains the importance of obtaining insight in class attendance, presents background information on the StudentLife dataset, and briefly introduces the method of building a predictive model for class attendance. The background on these topics is needed to grasp the scope of this thesis, and will be explained in more detail in following chapters.

## 1.1 Background

A student's life is not an easy life. It is expected of students to have a job because of financial concerns, have a social life, stay active for health purposes, and perform well academically (Wang et al., 2014). Furthermore, it is expected of students to attend classes provided by universities. Performing well is equivalent to receiving a high Grade Point Average (GPA) in the United States. According to previous research, GPA correlates with class attendance (Kassarnig, Bjerre-Nielsen, Mones, Lehmann, & Lassen, 2017). Obtaining insight in class attendance and predicting whether a student will or will not attend class is valuable information. Universities are able to optimize their educational program and to schedule classrooms using this information. Furthermore, professors are able to explain why a student does not perform well in class. Detecting low performers and intervening in time is not always possible for professors based on GPA. Professors are able to intervene earlier and prevent students from performing badly if they know students will or will not attend class. Instead of a reactive state of mind through evaluating students' GPA, a proactive state of mind can be created by increasing class attendance. Numerous studies, for instance research done by Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017) and Wang et al. (2014), focused on predicting students' performance. Little research has been done on predicting students' class attendance. Since class attendance is an important feature for predicting GPA, this feature should be investigated more extensively. Having knowledge of students' class attendance is interesting information for students, professors, and universities. Multiple studies, for instance research done by Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017), Nyamapfene (2010), and Chen and Lin (2008), indicated correlation between students' class attendance and students' performance. Therefore, if students do not attend class it is expected that they need to invest more time in studying at home to reach the same performance level as students that do attend class. In addition, absent students can fail to learn the most important information which is dealt with during a class and students will become less well educated and cause them study delay. In sum, predicting class attendance is not a topic that has been investigated extensively before. It is important to investigate the probability of predicting class attendance for students' personal development, for preparation and intervention of professors, and for optimizing universities' educational program.

In previous studies, predicting students' attendance was executed by making use of data-driven research. According to Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017), the applied methodology has limitations. It was stated that "results are based on analyzing surveys, sign-in-sheets or other types of self-reports, which are known to be prone to bias and errors". Meaning that students could respond 'attending class', when they did not attend class in real life. Eagle, Pentland, and Lazer (2009) stated that collecting data from mobile phones has 'the potential to provide insight into the underlying relational dynamics of organizations, communities and, potentially, societies'. Furthermore, it was specified that obtaining class attendance via surveys is biased and by analyzing mobile data, privacy issues have to be considered. Therefore, this research will focus on predicting class attendance based on mobile sensor data. A positive note on using mobile sensor data is that data is obtained without spending money (Eagle, Pentland, & Lazer, 2009). A comparable study is the StudentLife study of which data has been used to predict students' GPA based on, among other features, class attendance. The dataset provided by the StudentLife study will be used in this thesis as input for building different models to predict class attendance. Overall, because of bias created by students confirming they were attending class when they were not this study will only focus on mobile sensor data and education data obtained by the StudentLife dataset. Next, information on this dataset and further explanation on mobile sensor data will be presented shortly.

Students have a lot of everyday behavioral patterns and attending class is one of them. In this thesis, time series analysis of behavioral states derived from the StudentLife dataset (Wang et al., 2014) will be used. The StudentLife dataset contains data obtained from the phones of 48 undergraduate Dartmouth students who were tracked for a period of ten weeks. Data was obtained using four types of data gathering; sensor data, EMA data, pre and post survey responses, and educational data. Computational methods and Machine Learning algorithms on the phone were used to obtain behavioral trends (i.e., sleep, sociability, activity, stress, etc.). As stated before, in previous research building predictive models for students' performance was executed by making use of data-driven research of which the "results are based on analyzing surveys, sign-in-sheets or other types of self-reports, which are known to be prone to bias and errors", Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017). Therefore, this research will only focus on sensor data, consisting of GPS Location data and WiFi Location data (i.e., Students' Latitude, Students' Longitude, and Students' Location) and education data, consisting of Class Information data (i.e., Class Name, Class Start Time, Class End Time, and Class Location). This information is displayed in Table 1.1.

| StudentLife data | Predictors | Features |
|---|---|---|
| Sensor data | GPS Location data | Students' Latitude &<br>Students' Longitude |
| | WiFi Location data | Students' Location (i.e., building name) |
| Education data | Class Information data | Class Name, Class Start Time, End Time Class, and<br>Class Location (i.e., building name) |

Table 1.1: *Explanation of features used for predicting class attendance.*

By combining features displayed in Table 1.1, knowledge of students attending or not attending class will be obtained. Analyzing this knowledge and training the data enables to predict students' class attendance based on different predictors and on different moments in time prior to the class. Four experiments predicting class attendance from sensor data and education data obtained via smartphones will be proposed later in this research. In sum, this study focuses on building a predictive model based on Class Information data, GPS Location data, and WiFi Location data based on data collected of 48 Dartmouth students by researchers of the StudentLife study.

This paragraph first showed that class attendance is an interesting topic to investigate more extensively, then explained why mobile sensor data will be used to predict class attendance, and finally shortly provided detailed information on the StudentLife study. In addition, predicting class attendance based on mobile sensor data is assumed to avoid bias. It would be an opportunity to see if previous predictive models can be improved to predict class attendance. The insights gained on these topics will be used to setup a specific problem statement in the following paragraph.

### 1.2 Problem Statement

The previous paragraph described that predicting class attendance might have a positive impact on the quality of the education program. In this paragraph, this general observation will be turned into a concrete problem statement.

This thesis will focus purely on predicting class attendance. As described in the previous paragraph, having knowledge of students' class attendance is valuable information for students, professors, and universities. Improving predictive models can even be more valuable taking into consideration that class attendance can improve students' performance. When professors are able to detect performance issues related to class attendance in an early stage, that information can be exploited in the future to determine what type of student attends class regularly and will attend class in the future.

There are multiple ways of building predictive models. Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017) used Spearman's correlation coefficient to measure the correlation between

variables. Furthermore, they compared the distribution of variables by using Kruskal-Wallis H-test. The researchers used Dunn's multiple comparison tests with Bonferroni correction as follow-up post-hoc test. Moreover, they used Theil-Sen estimator to observe temporal trends in the dataset. Wang et al. (2014) used Pearson correlation analysis to understand the relationship between variables. In this thesis, machine learning techniques will be used for predicting class attendance. Predicting whether students will 'attend class' or 'not attend class', is a classification task. Therefore, building a classification algorithm is required. Possible techniques for predicting class attendance are the Logistic Regression algorithm, Naïve Bayes algorithm, and Random Forest algorithm, which all will be used in this thesis as approaches for predicting class attendance. These three models have differences in the complexity of predicting class attendance which will be explained in Section 2. In this thesis, sensor data and education data obtained by the StudentLife study are introduced into the field of pedagogy to investigate whether the prediction models can be further improved to predict class attendance. This will lead to the following problem statement:

*Can class attendance be predicted based on sensor data and education data?*

In this thesis, sensor data containing predictors GPS Location data and WiFi Location data, and education data containing predictor Class Information data will be used. These predictors and features, as displayed in Table 1.1, will be explained more extensively in Section 2. In sum, this paragraph proposed the problem statement and provided insights in previously used methods of research from comparable studies. In the next paragraph, it is explained how the problem statement will be answered.

### 1.3 Outline of the Thesis

The previous paragraph provided the problem statement. To answer the proposed question "*Can class attendance be predicted based on sensor data and education data"*, this thesis will be divided into four sub-questions. A general hypothesis before dividing the thesis is that all students follow the same courses. The problem statement will be answered by investigating the following four sub-questions:

1. *How well can class attendance be predicted based on Class Information data?*

The first analysis is about the degree the predictor Class Information data has influence on the prediction accuracy of the experiment. Because Class Location is used for determining class attendance, this feature will not be used as predictor in the first experiment. Therefore, the features Class Name, Class Start Time, and Class End Time will be obtained and taken into consideration. These features will be used as predictors in the predictive model. The hypothesis is that Experiment 2 will have a higher accuracy score than Experiment 1, and therefore performs significantly better than Experiment 1.

2. *How well can class attendance be predicted based on GPS Location data?*

The second analysis is about the degree the predictor GPS Location data has influence on the prediction accuracy of the experiment. Students' Latitude and students' Longitude will be obtained and taken into consideration in Experiment 2. The difference between the first and second research question is that the second research question contains more valuable information for predicting class attendance. The first research question focused on Class Information data and does not contain information on students' location at a certain moment. The questions are similar because both focus on one predictor and assess the degree the predictor has influence on the prediction accuracy of the model. It is hypothesized that the prediction accuracy of the second research question will be higher than the prediction accuracy of the first research question. It is expected that Experiment 4 will have a higher accuracy score than Experiment 2 and therefore performs significantly better using Class Information data, GPS Location data, and WiFi Location data than the experiment using only Class Information data or GPS Location data for predicting class attendance.

3. *How well can class attendance be predicted based on GPS Location data and WiFi Location data obtained prior to the start of the class?*

The third analysis investigates the degree the predictor WiFi Location data has influence on the prediction accuracy of the experiment. The difference between the previously determined research questions is that the third experiment focuses on two predictors. With this analysis it can be determined how much better class attendance can be predicted based on two valuable predictors containing information on where the student is at a certain moment. This can then be compared against the results of the first two experiments. Previously, data was obtained between the start time of the class and the end time of the class. This will be explained more extensively in Section 2. However, in the third research question data is obtained prior to the start of the class. For that reason, WiFi Location data can be used as predictor. Experiment 3 will be divided into three parts to answer the third question and investigate the impact. The analyses contain GPS Location data and WiFi Location data obtained one hour, three hours, and six hours prior to the start of the class. Students' Latitude, students' Longitude, and students' WiFi Location obtained one hour, three hours, and six hours prior to the start of the class will be obtained and taken into consideration. It is hypothesized that Experiment 3 will have higher prediction accuracy than Experiment 1 and Experiment 2. Furthermore, it is expected that when the delta between the moment of prediction and the moment the class will start increases, the accuracy of predicting class attendance will decrease. Meaning that Experiment 3 containing GPS Location data and WiFi Location data based on one hour prior to the class will have a higher accuracy score and therefore performs significantly better on predicting class

attendance than the model using GPS Location data and WiFi Location data obtained three and six hours prior to the start of the class. Furthermore, it is expected that the model containing GPS Location data and WiFi Location data six hours prior to the start of the class will perform worse than the experiments using GPS Location data and WiFi Location data obtained one hour and three hours prior to the start of the class.

4. *How well can class attendance be predicted based on GPS Location data, WiFi Location data and Class Information data?*

The last analysis investigates the degree the three predictors together have influence on the prediction accuracy of the fourth experiment. The difference between the fourth research question and the previously determined research questions is that the best performing model of Experiment 3 is used in combination with Class Information data including the feature Class Location. To answer this question, the best performing model of Experiment 3 (i.e., GPS Location data and WiFi Location data obtained one hour, three hours, or six hours prior to the start of the class) will be used as predictor. The features Class Name, Class Start Time, Class End Time, Class Location, students' Latitude, students' Longitude, and students' WiFi Location will be obtained and taken into consideration. It is hypothesized that Experiment 4 will be overall the best performing experiment. It is expected that the experiment will have a higher accuracy score than Experiment 1, Experiment 2, and Experiment 3.

In sum, to answer the research question, this thesis is divided into four sub-questions which will be used in Section 3 for conducting four experiments for predicting class attendance. The results of the performance of the four experiments will be presented in Section 4. Furthermore, Section 5 will provide a general discussion, present the conclusions of this thesis, and provide recommendations for further research. Next, a literature study is presented on related work in this domain in Section 2.

# Section 2: Related Work

The previous section provided the problem statement of this thesis and proposed the research questions to answer the problem statement. This section provides related work that has been executed on predicting class attendance. Research by Wang et al. (2014), Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017), and Zhou et al. (2016) are highlighted. This section determines that predicting class attendance is not an issue that has been investigated extensively before and provides background information on the classifiers and techniques used in this thesis.

### 2.1 Research Issue

It is interesting to investigate the opportunity of predicting class attendance more extensively since many studies found class attendance being a significant predictor of students' performance. A study by Nyamapfene (2010) showed that "class attendance is highly correlated to academic performance, despite the availability of online class notes". In addition, Westerman, Perez-Batres, Coffey, and Pouder (2011) stated that their results "suggest an inverse relationship between absenteeism and performance in management courses". The researchers showed that the negative effects of absenteeism have more impact on low performers. The negative effects of absenteeism seem to have no significant effect on high performers in the class. Likewise, Chen and Lin (2008) determined class attendance having a significant and positive impact on students' exam performances. In sum, according to a variety of previous research class attendance is a significant predictor for students' performance. Negative effects will reflect on students' performance if students do not attend class. For that reason, it is important for professors to step in early to prevent students from performing badly on exams. Having knowledge about whether a student will or will not attend class helps professors to prevent students from failing the course. Help could come too late if professors wait for insight in students' performance measured by exams. Having knowledge in whether a student will or will not attend class helps universities to optimize the educational program and to create better insight in how to schedule classes. Universities can decide to join classes or to schedule smaller classrooms if it is known that students do not attend class. This research issue is a matter of pedagogy. In sum, many studies found a correlation between class attendance and students' GPA. To be able to help students in an early stage during the course and to optimize universities education program, class attendance is an important subject to investigate more extensively. The following paragraph will provide better insight in related work that has been executed on the subject of predicting class attendance or other scholarly attributes based on mobile sensor data.

## 2.2 Related Work

In the previous paragraph, it was mentioned that the research issue is a matter of pedagogy. Therefore, this paragraph focuses on related work that has been done in building predictive models based on mobile sensor data in the field of pedagogy.

"*How Smartphones Can Assess and Predict Academic Performance of College Students*", is the title of research executed by Wang et al. (2014). The researchers stated the following: "the SmartGPA study uses passive sensing data and self-reports from students' smartphones to understand individual behavioral differences between high and low performers during a single 10-week term". The first contribution that can be derived from this study is the proposition of new methods to automatically infer study and social behaviors by making use of passive sensing from smartphones. The second contribution is performing time series analysis on the information obtained in the StudentLife dataset to discover which behaviors significantly impact term and cumulative GPA. The third contribution consists of two new proposed behavioral metrics to understand changes in behavior across the term. Finally, the researchers introduced a model that predicts students' cumulative GPA for the first time. According to Wang et al. (2014), their research extends previous research "by building a predictive model of academic performance based on students' self-reports and sensed behavior features obtained from their smartphones". The researchers used "linear regression analyses with lasso regularization to identify non-redundant predictors among a large number of input features" in order to perform the predictive analyses. This for the reason that predicting GPA is a regression task. The predictors used to build the predictive model were "automatic sensing time series behavioral data (i.e., conversational a study features), EMA time series data (e.g., positive affect and stress), mental health data (i.e., depression), and personality data (i.e., conscientiousness)". The researchers evaluated various regression models by using regularized linear regression, regression trees, and support vector regression with cross-validation to build the predictive model. The researchers have chosen to select the Lasso regularized linear regression model as their predictive model. Wang et al. (2014) stated that "Lasso automatically selects more relevant features and discards redundant features to avoid overfitting". The researchers used the mean absolute error (MAE), the coefficient of determination (*R*-squared), and Pearson correlation to evaluate how well the model performed. The researchers applied leave-one-subject-out cross validation to determine what the optimal parameters for Lasso were and determined the weights per feature. The study found among other things that study duration was a significant predictor of performance and that class attendance was not a significant predictor of students' performance. Wang et al. (2014) believe that "students' attendance is determined by the classes they take. Since all of them take at least one programming class, high achievers may not need to attend lectures to perform well". The results of the predictive model were that "GPA strongly correlates with the ground truth with $r = 0.81$ and $p < 0.001$, MAE is 0.179; *R*-squared is 0.559".

According to Wang et al. (2014), their "prediction model indicated that students receiving better grades are more conscientious, study more, experience lower level of stress as the term progresses, are less social in terms of conversations during the evening, and experience change in their conversation duration pattern later in the term". The researchers pointed out some limitations of their work. For instance, the dataset is large where the number of students in the study is small (N=30). Therefore, the small dataset limited the researchers to use more sophisticated models because those models might lead to overfitting. Furthermore, Dartmouth is an Ivy League school and the undergraduate students are the top performers of their high schools. Therefore, the sample is skewed whereby the students are high performers with good GPAs. The last limitation is that the students in the sample were not all computer science majors. However, all the students in the sample took one class they all had in common (i.e., Android programming). "The samples therefore could be biased to science students and do not represent the larger cross section of students found in liberal arts", Wang et al. (2014). In addition, the researchers use "location, date (i.e., weekday M-F) and time to automatically determine if a student attends a class or not". This is done by checking whether dwell time at the class location at least equals 90% of the scheduled period (e.g., 110 minutes). Making use of this approach allows the phone to automatically determine the class a student is taking and their attendance rates.

As stated before, Wang et al. (2014) did not find class attendance being a significant predictor of students' performance. A study that did found class attendance being a significant predictor for students' performance is the study by Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017). According to the researchers, a variety of data-driven research has been conducted on class attendance, absenteeism and the impact on students' performance. However, the applied methodology in previous research has limitations. It was stated that "results are based on analyzing surveys, sign-in-sheets or other types of self-reports, which are known to be prone to bias and errors". Therefore, the aim of the study by Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017) was "to evaluate the accuracy of measuring class attendance from smartphone data and assess its usefulness for discovering new patterns in data". The researchers used data collected by the Copenhagen Networks study who gathered data of the academic years 2013/14 and 2014/15. Data was collected using smartphones of nearly 1000 undergraduate students of the Technical University of Denmark. They kept track of the students' location by GPS, proximity of other students by Bluetooth, and mobile phone communication of the student. Only courses with the length of four hours containing at least eight students were considered in their research. The researchers had to calculate the location of the classes before they could determine class attendance of each student to work with the locations of the classes. The attendance was based on the location of the student relative to the estimated class location in each bin. Students were assigned to attend class when they were less than 200 meters away from the estimated class location and when this was measured in three different time

bins. Using the Bluetooth scans, the researchers calculated nearby students being within a distance of 15 meters. Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017) made use of Spearman's correlation coefficient to measure the correlation between the variables. This coefficient does not assume a linear relationship between two variables. The distributions of the variables were compared by using the Kruskal-Wallis H-test. In their study, Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017) showed that "attendance is correlated with the achieved grades both at the level of a specific course and overall performance (i.e., average term grade)". Furthermore, it was shown that "there is a general decrease in attendance over the course of a semester regardless of the performance. However, the attendance behavior of low and high performing students displays substantial differences over time". Lastly, the researchers showed the extent to which students have similar attendance patterns as their social peers. Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017) pointed out some limitations of their work. For instance, "estimation of class locations is based on Bluetooth and GPS signals, both of which are subject to noise", (Kassarnig, Bjerre-Nielsen, Mones, Lehmann, & Lassen, 2017). Furthermore, students who participated in the study were not the average student. They differ from the average students as they perform better. The limitation of the study is that the researchers only have measurements of a subset of the students following classes.

Other than the StudentLife study and the study by Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017), the study by Zhou et al. (2016) did not predict students' performance. Zhou et al. (2016) proposed "the EDUM (i.e., EDUcation Measurement) system to help characterize educational behavior through data collected from WLANs (i.e., WiFi networks) on campuses". One of the characterizations of educational behavior was class attendance. According to the researchers, EDUM uses longitudinal WLAN data to obtain insight in students' punctuality (i.e., students' attendance, students' arrival after start of the class, and students' early departures). Zhou et al. (2016) obtained data from 700 students measured for a period of 9 weeks at the Tsinghua University. Contribution of the study is to "design a scalable, non-intrusive, extensible and easy-to-deploy classroom education measurement system EDUM", Zhou et al. (2016). According to the researchers, they have chosen to adopt a relatively simple algorithm to infer whether a student is attending class or not. It was stated that to measure whether a student is attending class they needed to know whether the device of the student was close to the venue of the class. These observations were both measured by the WiFi data at the scheduled time of the class. Furthermore, the researchers had various assumptions about the data. For instance, a student could only be at one place at a time, every course is given at the same location and is rarely changed, the group of students attending class is relatively stable, and the recurring pattern of students regularly return to the course location corresponds with the schedule of the class. The attendance ratio of classes and timeslots is calculated by Zhou et al. (2016) by dividing the number of attended students by the number of appeared

students on campus. In addition, a second approach to calculate the attendance ratio was proposed by the researchers by dividing the number of classes attended per student by the number of classes the student appeared on campus. Furthermore, the researchers controlled for students arriving late and leaving early in class by dividing the number of late arrived students by the number of attended students and dividing the number of early escaped students by the number of attended students. Zhou et al. (2016) concluded that students with higher GPA attend classes more. Furthermore, it was stated that "EDUM is scalable, non-intrusive and extensible for new types of data and measurements", (Zhou et al., 2016). According to the researchers, limitations of the research are that students might turn their WiFi off during class and the researchers stated that their "metrics currently lack evaluations from manually collected data", (Zhou et al., 2016).

In sum, this thesis focuses on predicting class attendance. Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017) stated that the method of obtaining students' behavior by self-reports, done by most related work, ensures bias and errors. For that reason, this thesis will only include features of the StudentLife study that are not obtained by self-reports. This contains the GPS Location of the students and the Class Information data. Furthermore, the method proposed by Zhou et al. (2016) of determining whether a student is attending class, will be adopted in this thesis. Students' class attendance will be determined by investigating whether the WiFi Location of the student is equal to the scheduled WiFi Location of the classroom. Lastly, the method proposed by Wang et al. (2014) of using location, data i.e., weekday M-F and time to automatically determine if a student attends a class or not will be adopted in this thesis. This will be explained in more detail in Section 3.

### 2.3 The Classifiers and Techniques Used

Research questions to obtain insight in students' class attendance, to improve predictive models in order to improve students' performance and to detect performance issues related to class attendance in an early stage were proposed in Section 1. This paragraph shortly provides background information on the classifiers and techniques used to answer the proposed questions.

#### 2.3.1 Majority Baseline

The first classification technique considered to build a predictive model for class attendance is the Majority Baseline algorithm. It predicts class attendance based on the most frequent class in the data. For that reason, the Majority Baseline algorithm will either always predict correctly that students are not attending class and will never predict correctly that students are attending class, or the other way around. Meaning that the Majority Baseline algorithm is not a complex algorithm as it does not use a sophisticated formula for predicting the class based on unseen data. The performance of the Logistic Regression, Naïve Bayes, and Random Forest algorithms will be compared against the Majority Baseline

algorithm. The Logistic Regression, Naïve Bayes, and Random Forest algorithms use more sophisticated formulas to calculate the class based on unseen data. Therefore, it is expected that the three algorithms will always perform better in predicting class attendance compared to the Majority Baseline algorithm.

### 2.3.2 Logistic Regression

The second classification technique considered to build a predictive model for class attendance is Logistic Regression, a linear classification model. In this model, the probabilities describing the possible outcome of a single example are modeled by the logistic function. This function models how the probability *p* might be affected by one or more variables. Davey, Aiken, Hayes, and Hargreaves (2015) used the Logistic Regression algorithm to analyze health and educational impacts by examining "the association between the predictors' intervention and attendance". It was stated that their "analyses reflect attendance difference between treatment and control, but with modest statistical evidence". The Logistic Regression algorithm uses the logistic function to predict class based on unseen data. The algorithm uses a more sophisticated formula than the Majority Baseline algorithm to predict class attendance. Therefore, it is expected that the Logistic Regression algorithm will always perform better than the Majority Baseline algorithm. The usefulness of the Logistic Regression algorithm is that is uses the logistic function to calculate a number between 0 and 1 en rounds it to the nearest number, Le (2018). However, bias occurs when $p = 0.51$ is rounded up to '1'.

### 2.3.3 Naïve Bayes

The third classification technique considered to build a predictive model for class attendance is Naïve Bayes. This approach "works on the unrealistic assumption that (a) the contributions of all predictor variables to the overall prediction or classification are equally important, and (b) the effects of the predictors are independent of each another. These unrealistic assumptions, which give 'Naïve' Bayes its name, allow it to be quite computationally efficient, and to require very little training data for the development of parameter estimates", Attewell, Monaghan, and Kwong (2015). Furthermore, Attewell, Monaghan, and Kwong (2015) stated the following: "The probability of the outcome given the input(s) is the product of the probability of the outcome and the probability of the input(s) given the outcome, divided by the probability of the inputs". Research by Anuradha and Velmurugan (2016) investigated feature selecting techniques to analyze students' performance using Naïve Bayes classifier. The usefulness of Naïve Bayes algorithm is that it calculates the conditional probability of each class and assumes that the variables are independent and equally important, Le (2018). When this is calculated, the Bayes Theorem is used for predicting unseen data. It is expected that the Naïve Bayes algorithm will perform better than the Majority Baseline algorithm and the Logistic Regression algorithm because the Naïve Bayes algorithm uses a more sophisticated formula to predict class attendance.

### 2.3.4 Random Forest

The fourth classification technique considered to build a predictive model for class attendance is Random Forest. This approach is used to average the results of multiple decision trees to obtain the best prediction. "The novel aspect of Random Forests is that the researcher forces a different subset of predictors to be included in each model, so that each model cannot have an identical structure or content to the previous one. The varied predictions obtained from those multiple models are then combined to yield a best estimate", Attewell, Monaghan, and Kwong (2015). Research by Mythili and Shanavas (2014) used the Random Forest algorithm to predict students' performance using among others class attendance as predictor. The researchers stated the following: "it is discovered that Random Forest performance is best than that of different algorithms employed in the study". This algorithm is the most complex of all proposed algorithms and is a powerful machine learning algorithm. The Random Forest algorithm is useful because it combines the prediction results of each sample of the data. This results in a better estimate of the true underlying output value, Le (2018). The Random Forest algorithm uses a more complex formula to predict class attendance than the Majority Baseline, the Logistic Regression, and the Naïve Bayes algorithms. Therefore, it is expected that the Random Forest algorithm performs better than these three algorithms.

### 2.3.5 Cross-validation and overfitting

In addition, the research conducted by Anuradha and Velmurugan (2016) also used cross validation in their study. Cross-validation is fundamental for avoiding overfitting. "Some applications are too effective at building a predictive model; they construct something too complicated that will not generalize to other examples", Attewell, Monaghan, and Kwong (2015). If the model is trying to fit the data points too well, this is called overfitting. If overfitting occurs, the model does not only fit the data points, it is also fitting the noise. In this thesis, the k-fold cross-validation method will be used for validating the performance of the predictive model. Using k-fold cross-validation, the researcher is able to change the number of *k* manually. The number of *k* is randomly selecting subsamples. One of the *k* subsamples will initially function as the validation dataset to validate the performance of the training data, according to Attewell, Monaghan, and Kwong (2015).

### 2.3.6 Confusion Matrix

Research by Mueen, Zafar, and Manzoor (2016) investigated the possibility of predicting students' performance using data mining techniques. In their study, they made use of confusion matrices to measure fit of predictions of the model. According to Attewell, Monaghan, and Kwong (2015), "the confusion matrix informs us how accurately the predictive model we have constructed performs in classifying cases. It compared the predicted outcome (yes/no) with the observed or actual outcome

(yes/no)". The confusion matrix consists of four possible outcomes: true negative (TN), false positive (FP), false negative (FN), and true positive (TP). Correctly predicted observations are the following: those cases that were predicted as negative and were actually observed as negative (true negative) plus those that were predicted as positive and were observed as positive (true positive). For an accurate model most of its cases should ideally appear as true negative and true positive observations (Attewell, Monaghan, & Kwong, 2015). Furthermore, the cases that were predicted as positive and were actually observed as negative are called false positives. The cases that were predicted as negative and were observed as positive are called false negatives. Table 2.1 displays the general confusion matrix.

| **Confusion Matrix** | **Predicted not Attending** | **Predicted Attending** |
|---|---|---|
| Actual not Attending | TN | FP |
| Actual Attending | FN | TP |

Table 2.1: *Confusion Matrix explanation.*

In sum, this paragraph provided background information on the classifiers (i.e., Logistic Regression, Naïve Bayes, and Random Forest algorithms) and techniques (i.e., confusion matrix and cross-validation) used to answer the proposed questions.

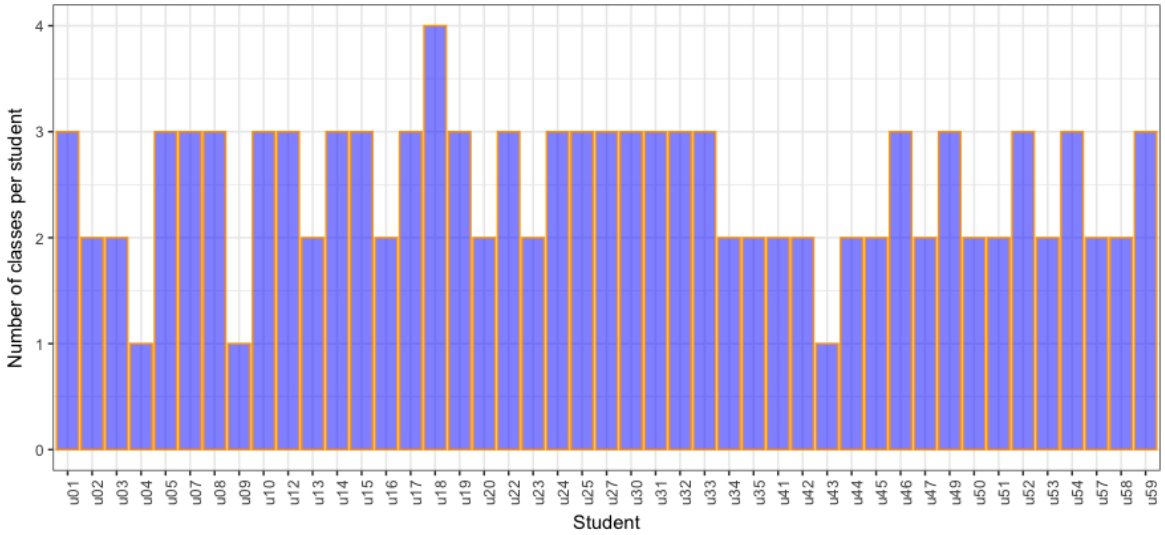
Overall, it was concluded that previous research focuses predominantly on predicting students' performance. Furthermore, students' performance correlated with class attendance in most of the previous research. It is therefore an interesting research subject to analyze and predict students' class attendance. Furthermore, this paragraph outlined studies by Wang et al. (2014), Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017), and Zhou et al. (2016) and discussed the contributions and limitations of their studies. This thesis will differ from the study by Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017) in the technique used to predict class attendance. This thesis will focus on machine learning techniques such as Logistic Regression algorithm, Random Forest algorithm, Naïve Bayes algorithm, accuracy score and F1-score. The study by Kassarnig, Bjerre-Nielsen, Mones, Lehmann, and Lassen (2017) focused on Kruskal-Wallis H-test and Spearmans' Spearman's correlation. Furthermore, this thesis will differ from the study by Zhou et al. (2016) by predicting class attendance based on WiFi Location data, GPS Location data, and Class Information data to predict class attendance. Zhou et al. (2016) used WiFi Location data alone for analyzing and determining class attendance. It will be discussed how the prediction of class attendance is executed in the method section of this thesis.

# Section 3: Method

Previously, related work of the classifiers was provided and the aim to investigate the probability of predicting class attendance for students' personal development, for preparation and intervention of professors, and for optimizing universities' educational program was explained. The following section will describe the reason that class attendance is measured by comparing the name of the building the class is given in to the name of the building the student is in or near. In addition, it will explain the classifiers and algorithms used to build the predictive model.

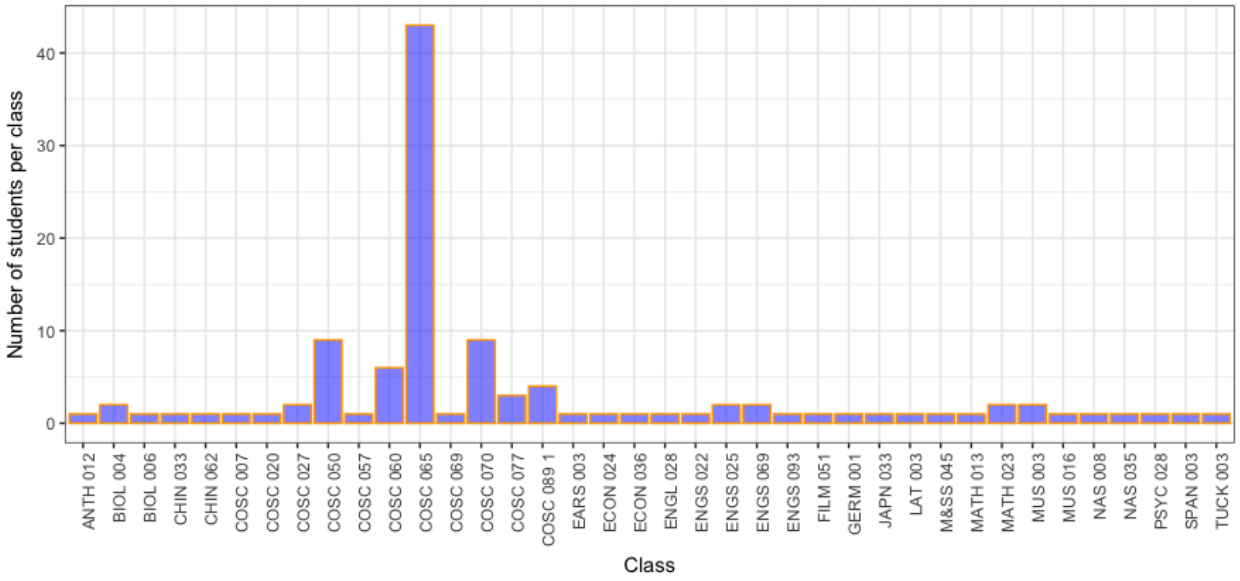
### 3.1 Dataset Information and Collection Methodology

As mentioned in previous sections, the StudentLife dataset was built during a study at Dartmouth University by Wang et al. (2014). The Dartmouth Universities' students have identification numbers varying from u01 to u59. However the number of students goes up to u59 only 48 students participated in the study. As can be seen in Graph 3.1, the dataset provided by the StudentLife study shows that most students (N=23) followed three courses in the spring term of 2013. Furthermore, three students who participated in the StudentLife study did not follow any courses (e.g., student u36, u39, u56). These three students are missing from Graph 3.1. This graph conflicts with the paper provided by Wang et al. (2014) since it was stated that 'each student takes three classes'. The following paragraph will explain that this problem is not the only limitation of the dataset provided by the StudentLife study. Further information on the number of classes followed per student provided by the StudentLife dataset shows different results. This information is displayed in Table A2 of the appendix.




Graph 3.1: *Number of classes followed per student.*

In addition, Graph 3.2 displays the number of students per class. It is concluded from this graph that almost all students follow the course called “COSC 065”. In the StudentLife study the researchers use abbreviations for the courses students are able to follow. It was not specified where these abbreviations stand for. However, it was stated that “all of them take at least one programming class”. For that reason it can be assumed that “COSC 065” is the programming course called “Android programming”. Table A1 in the Appendix shows that student u08, u32, and u51 attend most classes, respectively 22, 33, and 22 times.




Graph 3.2: *Number of students per class.*

It was stated that the number of classes taken by students represented in the by StudentLife provided dataset do not correspond with the statement of Wang et al. (2014) that ‘each student takes three classes’. This is not the only limitation of the StudentLife dataset. Firstly, the students in the StudentLife study could follow classes the dataset not provided information on. In the Appendix, Table A2 shows the information of the classes followed per student provided by the StudentLife dataset. Furthermore, Table A3 in the Appendix shows the information on the classes the students could follow at Dartmouth University provided by the StudentLife dataset. Table A2 shows that two students take class “COSC 098” but there is no general class information available of the course. Therefore, the dataset used in this thesis contains fewer observations than specified by the StudentLife study. Secondly, the GPS Location of the classes is missing from the dataset. The researchers of the StudentLife dataset changed the GPS Locations to meaningful names by labelling the locations with the name of the building the class was given in. Therefore, class attendance cannot be measured by calculating the distance of the student to the class based on GPS Location data. Furthermore, the researchers of the StudentLife dataset transformed the WiFi Location in numbers to meaningful names by labelling the WiFi Location data with the name of the

building the student was in or near. Therefore, class attendance will be measured by comparing the name of the building the class is given in to the name of the building the student is in or near.

As determined in Section 2, class attendance was in previous research defined based on location, date, and time (Wang et al., 2014). The examples were students being less than 200 meters away from the estimated class location when this was measured in three different time bins (Kassarnig, Bjerre-Nielsen, Mones, Lehmann, & Lassen, 2017), dividing the number of attended students by the number of appeared students on campus, and dividing the number of classes attended per student by the number of classes the student appeared on campus (Zhou et al., 2016). As stated in the previous paragraph, the researchers of the StudentLife study labeled the locations of the buildings with meaningful names. Therefore, the approach of measuring students' distance from the estimated class location will not be possible using the StudentLife dataset. In addition, dividing the number of attended students by the number of appeared students and dividing the number of classes attended per student by the number of classes the student appeared on campus cannot be calculated using the StudentLife dataset. This thesis focuses on the GPS Location and WiFi Location of Dartmouth Universities' students and does not contain real life information on the number of students attending certain classes. The following paragraph will explain more extensively how the name of the building the class is given in will be compared to the name of the building the student is in or near to determine class attendance.

In this thesis, location, date, and time are used to determine whether a student attends class or not. As stated in Section 2, this method was also used by the researchers of the StudentLife study. To be more concrete, the WiFi Location and the Class Location of the student, the date the class is given, and the time between the start and end of the class are used to determine whether a student is attending class. The feature Class Location consists of the name of the building the class is given in. The feature WiFi Location consists of the name of the building the student is in or near. Class attendance will be determined by comparing whether the student was in or near the building the class is given in, based on the students' location measured between the start time and end time of the class. For instance, if the class started at 11:15 and ended at 12:20, all WiFi Location data of the student obtained between these two times will be used. If students' WiFi Location is at least once equivalent to the Class Location, class attendance is measured as '1'. This information is briefly illustrated in Table 3.1.

| Student | Date | Start Class | End Class | Location Class | WiFi Location | Attendance |
|---|---|---|---|---|---|---|
| u01 | 2013-03-29 | 11:15:00 | 12:20:00 | hopkins | reed | 0 |
| u01 | 2013-04-01 | 11:15:00 | 12:20:00 | lsb | lsb | 1 |
| u01 | 2013-04-04 | 11:15:00 | 12:20:00 | NA | kemeny | 0 |

Table 3.1: *Example explanation of determining students' class attendance.*

In sum, this paragraph provided additional information of the StudentLife dataset and provided insight in the number of classes followed per student and the number of attended classes per student, in the classes the Dartmouth University provides according to the StudentLife dataset, and how many students are following a course. This paragraph also explained that class attendance will be measured by comparing the name of the building the class is given in to the name of the building the student is in or near. The next paragraph will build further on this information.

### 3.2 Data Cleaning

The previous paragraph explained that class attendance will be determined based on the features Class Location and students' WiFi Location. This paragraph explains how the data is cleaned in order for the dataset to be useful for making predictions on.

Data cleaning is the process of detecting and correcting or removing inconsistent, inaccurate and noisy data from a dataset. The StudentLife dataset contains few missing values. The Listwise Deletion Approach is used to remove entire rows containing missing values whenever missing values occur. The StudentLife study provided information separated per subject per student in multiple datasets. Duplicates emerge when datasets are combined. These are removed from the dataset. Combining datasets will be explained in more detail in later paragraphs. Furthermore, data type of the features is transformed from a number to time by changing the timestamp. For instance, the number 0.4687500 becomes the timestamp 11:15:00. Figure 3.1 shows the process of selecting instances and features to create the subset of the dataset this thesis will work with.

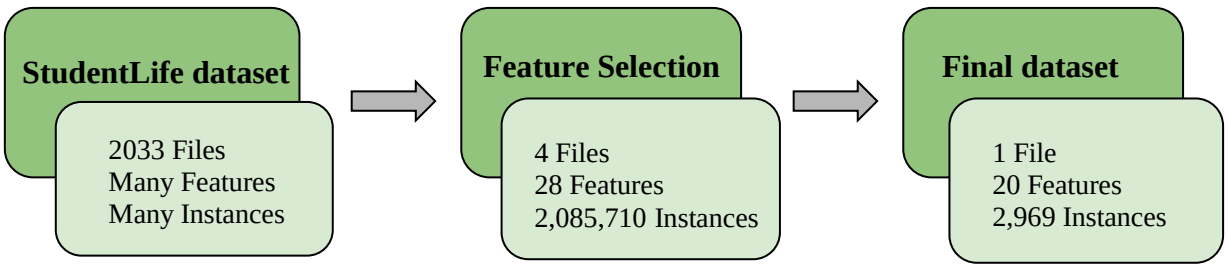



Figure 3.1: *Process of selecting instances and features for subset.*

In sum, this paragraph provided information on how the datasets are cleaned before different data files are combined into one dataset. The following paragraph will explain the feature selection of this process more extensively.

### 3.3 Feature Selection

The previous paragraph provided information on how the dataset was transformed so predictions could be made. The following paragraph is an addition to the previous paragraph. It explains the selected features and provides information on how these are transformed to build the predictive model on.

Feature selection means reducing the number of variables or features to be included in a model by identifying the important ones and dropping the rest, in such a way that those remaining can still accurately predict the target (Attewell, Monaghan, & Kwong, 2015). The aim is to identify which predictors are strongly associated with an outcome variable of interest, according to Attewell, Monaghan, and Kwong (2015). As described in the first paragraph of this section this study focuses on predicting class attendance based on sensor data (e.g., GPS Location data and WiFi Location data) and education data (e.g., Class Information data). Results based on analyzing surveys, sign-in sheets or other types of self-reports are prone to bias and errors (Kassarnig, Bjerre-Nielsen, Mones, Lehmann, & Lassen, 2017). Therefore, the predictors GPS Location data, WiFi Location data, and Class Information data are selected as input for this research. Data on these features are obtained without surveys, sign-in-sheets or other types of self-report. The remainder of this paragraph will explain the obtained features and what variables the features consist of.

#### 3.3.1 Class Information data

The feature Class Information contains general information on Dartmouth Universities' courses. The participants of the StudentLife study could follow 45 different courses. It was hypothesized that most students follow the same courses before doing the descriptive analyses. On the contrary, 26 of the 45 courses are taken by only one participant. The feature Class Information contains information on the Class Name, Class Start Time, Class End Time, and Class Location. To be able to process the Class Information data each row is multiplied ten times. This action is executed because GPS Location and WiFi Location are tracked for a period of ten weeks. The location of the classes was determined in the StudentLife study by the name of the building the classes were given in. The locations were labeled with meaningful names by the researchers. Examples of the names of the buildings are "lsb", "kemeny", or "silsby-rocky".

### 3.3.2 WiFi Location data

The WiFi Location dataset consists of students' Location and is predominantly tracked every five minutes. However, occasionally the data is recorded every twenty or even every two minutes. There is no consistency in the interval of obtaining the data. The researchers kept track of the time and students' Location to retrieve WiFi Location data. The WiFi Location data consists of the name of the building the student is in or near (i.e., in[building] or near[building]). Students' WiFi Location data is used to determine whether a student is attending class or not. The student is classified as 'attending class' if students' WiFi Location and Class Location are at least once equivalent between the Class Start Time and the Class End Time. The student is classified as 'not attending class' if students' WiFi Location and Class Location are not at any time equal between the start and end time of the class. Afterwards, duplicates are removed from the dataset to remain one observation per student, per date, per class. The time used to determine whether a student is attending class or not is therefore a random time between the start and the end of the class.

### 3.3.3 GPS Location data

The GPS Location data consists of students' Latitude and Longitude obtained per student. Predominantly, the GPS Location is obtained every twenty minutes. However, occasionally the GPS location is obtained every ten minutes. There is no consistency in the interval of obtaining the data. From this dataset the features Time, students' Latitude, and students' Longitude are subtracted in order to predict class attendance. The GPS Location data is combined to the students' WiFi Location data subset based on time. This is done based on the observation of GPS Location being the closest measurement to the random time the WiFi Location of the student is measured. For instance, if students' WiFi Location is measured at 11:20 the datasets are combined to the closest observation of the GPS Location data, for instance obtained at 11:21. In sum, a subset is created containing students' Latitude and students' Longitude observed at the time being the closest to the random WiFi Location data observation between the start and the end of the class.

### 3.3.4 GPS Location different moments in time

Previously, it was explained that GPS Location data (i.e., students' Latitude and students' Longitude) is obtained based on the measurement of the random WiFi Location data between the start and end of the class. However, it is interesting to investigate whether class attendance can be predicted based on different moments in time prior to the start of the class. This thesis will focus on the GPS Location data obtained one hour prior to the start of the class, three hours prior to the start of the class, and six hours prior to the start of the class. New features are created based on the GPS Location data of the measurement being closest to students' random WiFi Location observation between the start and end time

of the class minus one hour, three hours, or six hours. Meaning that six new features are created: (1) students Latitude minus one hour, (2) students' Longitude minus one hour, (3) students Latitude minus three hours, (4) students' Longitude minus three hours, (5) and students Latitude minus six hours, (6) students' Longitude minus six hours. The feature Attendance is determined as 'NA' because subscribing a 0 to them would give incorrect insights in whether a student is attending the class or not.

Overall, the reason the features GPS Location data, WiFi Location data, and Class Information data are selected is that these features can be obtained without inference of the students. This paragraph provided insight in how the features were obtained by the Studentlife study and explained how the different features are combined to one dataset. The next paragraph will explain how this dataset is prepared to build the predictive models on.

### 3.4 Data Preparation and Partitioning

The previous paragraph described how multiple datasets were transformed into one dataset. This paragraph will explain the alterations needed in order for the dataset to become suitable for building the predictive model on.

In the first instance, the dataset was unbalanced. Classification algorithms often simply categorize all cases as belonging to the majority class to minimize the error rate in predictions in the presence of highly unbalanced outcomes. The dataset was unbalanced because class attendance was measured based on one moment in time thirty minutes after the start of the class. This resulted in approximately 90% of the students not attending class and 10% of the students attending class. The new dataset is created based on all the WiFi Location observations between the start and the end time of the class. Therefore, there is a higher likelihood the student is measured at least once in the building the class is giving in between the start and end time of the class. Now, the dataset is balanced. There are 1645 unattended classes, and 1324 attended classes. In sum, this paragraph determined inconveniences of the dataset. It was described how these inconveniences were modified in order for the dataset to become suitable to build the predictive model on. The following paragraph will explain the number of attended class and unattended classes per experiment.

### 3.5 Experiments

In the previous paragraphs insights about students and the classes of the Dartmouth University were provided. Furthermore, it was explained how the different datasets are combined into one dataset appropriate for building a model to predict class attendance. This paragraph will explain how the four proposed sub-questions will be used to conduct four experiments. The sub-questions proposed in the

introduction will divide this section into four sub-sections containing information on what features are used to conduct each experiment. General modifications will be explained first.

To answer the research questions, several general modifications have to be conducted. The rows containing NA values are omitted from the dataset to answer the research questions. Furthermore, the dataset is split into test (20%) and training data (80%). In addition, "z-scores are a way to compare results from a test to a 'normal' population", Statistics How to (2018). By transforming the training and test set into z-scores, the files are enabled to be used in the proposed algorithms and to compare accuracy scores against the 'normal' population (i.e., the Majority Baseline). The Logistic Regression algorithm, Naïve Bayes algorithm, and Random Forest algorithm are used for building the predictive models on and the accuracy scores and the F1-scores of the algorithms are compared against the Majority Baseline. Results can be found in Section 4. Paragraph 3.6 will explain the accuracy score and F1-score more extensively.

3.5.1 Experiment 1: How well can class attendance be predicted based on Class Information data?

The first experiment for predicting class attendance is based on the following Class Information data features: Class Name, Class Start Time, and Class End Time. The feature Attendance is the target variable and consists of the number of 1645 unattended classes (0) and 1324 attended classes (1). The performance of the Majority Baseline algorithm is based on the most frequent class of the experiment. The most frequent class in Experiment 1 is the 'not attending' class (0). Therefore, it is expected that the Majority Baseline algorithm in Experiment 1 will always predict correctly that students are not attending class and will not ever predict correctly that students are attending class. It is expected that the Majority Baseline algorithm will perform equally well in Experiment 1 and Experiment 2 because the same selection of the data is used.

3.5.2 Experiment 2: How well can class attendance be predicted based on GPS Location data?

The second experiment for predicting class attendance is based on the following GPS Location data features: students' GPS Location Latitude and students' GPS Location Longitude. The feature Attendance is the target variable and consists of the number of 1645 unattended classes (0) and 1324 attended classes (1). As stated in the previous paragraph, the same selection of data is used in Experiment 1 and Experiment 2. Meaning that there will be no difference in the performance of the Majority Baseline algorithm. However, this experiment differs from Experiment 1 because it focusses on students' location data. For that reason, it is expected that the Logistic Regression, the Naïve Bayes, and the Random Forest algorithms perform better in predicting class attendance than the same algorithms in Experiment 1.

3.5.3 Experiment 3: How well can class attendance be predicted based on GPS Location data and WiFi Location data obtained prior to the start of the class?

The third experiment for predicting class attendance is based on the following features: students' GPS Location Latitude obtained prior to the start of the class, students' GPS Location Longitude obtained prior to the start of the class, and students' WiFi Location obtained prior to the start of the class. The feature Attendance is the target variable based on data obtained one hour prior to the start of the class and consists of the number of 1551 unattended classes (0) and 1281 attended classes (1). The feature Attendance based on data obtained three hours prior to the start of the class consists of the number of 1524 unattended classes (0) and 1246 attended classes (1). The feature Attendance based on data obtained six hours prior to the start of the class consists of the number of 1517 unattended classes (0) and 1250 attended classes (1). There are three different selections of the dataset used for predicting class attendance. Therefore, Experiment 3 consists of three different values of the Majority Baseline algorithm. The most frequent class is the 'not attending' class (0) in all three the situations. However, the different numbers of observations (i.e., respectively 1551, 1524, and 1517) result in different performance scores of the Majority Baseline algorithm. Experiment 3 is similar to Experiment 1 and Experiment 2 because the same algorithms are used for predicting class attendance. However, Experiment 3 differs from Experiment 1 and Experiment 2 because it uses three different samples of the dataset. These different samples are obtained at different moments in time prior to the start of the class. This will expand the investigation of predicting class attendance because data that is not obtained at the same time attendance is determined. In addition, class attendance is determined based on Class Location and WiFi Location. Meaning that WiFi Location data can be used in Experiment 3 because data is obtained prior to the moment in time that class attendance was determined. It is expected that the Logistic Regression algorithm, the Naïve Bayes algorithm, and the Random Forest algorithm perform better when the WiFi Location data is added as predictor compared to Experiment 1 and Experiment 2.

### 3.5.4 Experiment 4: How well can class attendance be predicted based on GPS Location data, WiFi Location data, and Class Information data?

The fourth experiment for predicting class attendance is based on the following features: Class Name, Class Start Time, Class End Time, Class Location, students' GPS Location Latitude obtained prior to the start of the class, students' GPS Location Longitude obtained prior to the start of the class, and students' WiFi Location obtained prior to the start of the class. The feature Attendance is the target variable and is based on data obtained prior to the start of the class. The target variable consists of the same number unattended classes (0) and attended classes (1) as data obtained one hour, three hours, or six hours prior to the start of the class proposed in Experiment 3. As determined in Section 1, the best performing experiment of Experiment 3 is used to answer the last research question. Meaning that there will be no difference in the performance of the Majority Baseline algorithm in Experiment 4 and the best performing

experiment in Experiment 3. The last experiment differs from the other experiments because the best performing selection of the data determined in Experiment 3 is used for predicting class attendance. Furthermore, the feature Class Information is added as predictor. As stated in the previous paragraph, WiFi Location data could be used in Experiment 3 because data is used that is obtained prior to the moment in time that class attendance was determined. For that same reason Class Location is added as input variable to answer the fourth research question. It is expected that the Logistic Regression algorithm, the Naïve Bayes algorithm, and the Random Forest algorithm perform better when the WiFi Location data and the Class Location data are added as predictors. The number of observations per experiment and a brief summary of the predictors used per experiment for determining class attendance are shown in Table 3.2.

| **Class** | Students not attending (0) | Students attending (1) | Predictors per experiment |
|---|---|---|---|
| Experiment 1 | 1645 | 1324 | Class Name, Class Start Time, Class End Time |
| Experiment 2 | 1645 | 1324 | GPS Location Latitude, GPS Location Longitude |
| Experiment 3 (-1) | 1551 | 1281 | GPS Location Latitude -1 hour, GPS Location Longitude -1 hour, WiFi Location -1 hour |
| Experiment 3 (-3) | 1524 | 1246 | GPS Location Latitude -3 hours, GPS Location Longitude -3 hours, WiFi Location -3 hours |
| Experiment 3 (-6) | 1517 | 1250 | GPS Location Latitude -6 hours, GPS Location Longitude -6 hours, WiFi Location -6 hours |
| Experiment 4 | *Best performing model Experiment 3 is not Determined yet* | | Class Name, Class Start Time, Class End Time, Class Location, GPS Location Latitude -1, -3 or -6 hours, GPS Location Longitude -1, -3 or -6 hours, WiFi Location -1, -3 or -6 hours |

Table 3.2: *Number of students not attending class and attending class per experiment.*

In sum, this paragraph provided the features used per experiment and described the alterations needed per experiment in order for the dataset to become suitable for building the predictive model on. Furthermore, it provided the differences and similarities between the experiments. Next, the evaluation method will be explained.

### 3.6 Evaluation Method

The previous paragraphs provided insight in the classifiers used for predicting class attendance in different experiments. This paragraph explains the evaluation of the four proposed experiments. It explains how the accuracy and F1-scores are used to evaluate the performance of the experiments and algorithms.

The previously defined classifiers are evaluated by the classification accuracy score and the F1-score. The classification accuracy score is the proportion of instances which are correctly classified by the algorithm (Attewell, Monaghan, & Kwong, 2015). The F1-score is the harmonic mean of the precision and recall score. Precision calculates the true positives of the algorithm by dividing the true positives by the true positives plus the false positives. Recall calculates how often the algorithm captures the actual positives as positives by dividing the true positives by the true positives plus the false negatives (Attewell, Monaghan, & Kwong, 2015). The classification accuracy score and the F1-score are calculated using the Logistic Regression algorithm, the Naïve Bayes algorithm, and the Random Forest algorithm. These scores are compared against the Majority Baseline algorithm to evaluate how well the different algorithms perform on the data. The accuracy and F1-scores are values between 0 and 1. The closer the value of the scores is to 1, the better the model performs. However, if the accuracy and F1-scores perform better on the training set than on the test set, this might be an indication of overfitting. The accuracy score and F1-score of the Majority Baseline algorithm are penalized per model to retrieve a valid accuracy score and F1-score to compare the other algorithms against. The accuracy score is more useful when the false positive and false negative values are similar. If the false positives and false negatives are not similar it is better to look at the F1-score, Joshi (2016). In sum, the performance of the experiments and algorithms are compared using the accuracy score and the F1-score. The closer the scores are to 1, the better the performance of the algorithms is.

### 3.7 Software

To perform the analyses and build the predictive model the programs R (version 3.4.2) in RStudio and Python (version 3.3) are used. In R the following packages are used: rjson, ggplot2, dplyr, DT, jsonlite, magrittr, glmnet, data.table, plyr, chron, lubridate, tidyr, reshape2, and ggmap. The predictive model is built in Python. In Python the following packages are used: numpy and sklearn.

# Section 4: Experiments and Results

In the previous sections, the aim of this thesis was defined as investigating the probability of predicting class attendance for students' personal development, for preparation and intervention of professors, and to optimize universities' educational program. This section reports classification performance based on accuracy scores and F1-scores of the four experiments proposed in Section 3.

As described in Section 3, the dataset was split into a test set (20%) and training set (80%). K-fold cross-validation was applied to validate the training set. Table 4.0 displays the best penalty parameter value, the mean value, minimum value, and maximum value of the k-fold cross-validation per model. The table represents the cross-validation scores based on the four proposed experiments.

| | Best penalty Parameter value | Mean | Min | Max |
|---|---|---|---|---|
| Experiment 1 | C=1 | 0.56 | 0.56 | 0.56 |
| Experiment 2 | C=1 | 0.56 | 0.56 | 0.56 |
| Experiment 3 (-1) | C=1 | 0.65 | 0.65 | 0.65 |
| Experiment 3 (-3) | C=1 | 0.56 | 0.56 | 0.56 |
| Experiment 3 (-6) | C=1 | 0.55 | 0.55 | 0.55 |
| Experiment 4 | C=5 | 0.71 | 0.70 | 0.71 |

Table 4.0: *Cross-validation values of predicting students' class attendance based on different experiments.*

As can be seen in Table 4.0, there is variation in the cross-validation score of the experiments. Important values of Table 4.0 are the best penalty parameter value and the mean value. The mean value is used to compare the performance of the experiments. The minimum and maximum cross-validation scores are reported to better evaluate the cross-validation score per experiment. As can be seen in the Table 4.0, there is not much variance between the minimum and the maximum values. Meaning that the cross-validation scores of the experiments will not differ significantly when another penalty parameter value is chosen. Table 4.0 shows that Experiment 1 and Experiment 2 perform equally well, a result that contradicts the expectations. Table 4.0 shows that Experiment 3 containing data obtained one hour prior to the start of the class performs the best of the three experiments, a result that supports the expectations. In addition, the experiment using data obtained three hours prior to the start of the class performs better than the experiment using data obtained six hours prior to the start of the class, a result that supports the expectations. Lastly, Experiment 4 is the best performing model, a result that supports the expectations. In sum, inconsistency of the agreement of expectations derives based on the cross-validation scores. Conclusions based on the observations can be found in Section 5. The following paragraphs provide results on the performance of the test set per experiment based on accuracy scores and F1-scores. The

classifiers Logistic Regression, Naïve Bayes, and Random Forest were applied in all experiments and compared against the Majority Baseline algorithm in order to evaluate the performance per algorithm. The accuracy scores and F1-scores of the test set were computed per classifier per experiment as a measure for comparison between the classification performances of the models.

### 4.1 Results Experiment 1: How well can class attendance be predicted based on Class Information data?

The features selected for analyzing the classification performance of the models for predicting class attendance based on Class Information data were: Class Name, Class Start Time, and Class End Time. The target variable was Attendance. The accuracy scores and F1-scores of the first experiment are shown in Table 4.1.1.

| **Classifier** | Accuracy (training) | Accuracy (test) | F1-score (training) | F1-score (test) |
|---|---|---|---|---|
| Majority Baseline | 0.56 | 0.53 | 0.72 | 0.69 |
| Logistic Regression | 0.58 | 0.55 | 0.70 | 0.68 |
| Naïve Bayes | 0.56 | 0.53 | 0.72 | 0.69 |
| Random Forest | 0.68 | 0.68 | 0.69 | 0.69 |

Table 4.1.1: *Performance accuracy scores and F1-scores of predicting students' class attendance based on Class Information data.*

Based on the performance scores per algorithm, Table 4.1.1 displays the Logistic Regression algorithm and the Random Forest algorithm performing better than the Majority Baseline algorithm. The Naïve Bayes algorithm performed equal to the Majority Baseline algorithm, a result that contradicts the expectations. Furthermore, the Random Forest algorithm performed better than the Logistic Regression and the Naïve Bayes algorithms, a result that supports the expectations. However, the Logistic Regression algorithm performed better than the Naïve Bayes algorithm, a result that contradicts the expectations. In addition, none of the models were overfitting since the scores of the training set were approximately equal to the scores of the test set. The accuracy score of the Random Forest algorithm showed a better performance than the Logistic Regression algorithm and the Naïve Bayes algorithm where the F1-score showed the three algorithms performing equally well, a result that contradicts the expectations. Therefore, according to Table 4.1.1, the approach of Random Forest algorithm yields the best classification performance. The confusion matrices per model of Experiment 1 per algorithm will be provided next.

| | Predicted not Attending | Predicted Attending |
|---|---|---|
| Majority Baseline | | |
| Actual not Attending | 316 | 0 |
| Actual Attending | 278 | 0 |
| Logistic Regression | | |
| Actual not Attending | 310 | 6 |
| Actual Attending | 259 | 19 |
| Naïve Bayes | | |
| Actual not Attending | 316 | 0 |
| Actual Attending | 278 | 0 |
| Random Forest | | |
| Actual not Attending | 154 | 162 |
| Actual Attending | 31 | 247 |

Table 4.1.2 *Test set Confusion Matrix values per algorithm based on Class Information data.*

Table 4.1.2 displays the confusion matrices per algorithm for Experiment 1. Firstly, the Logistic Regression algorithm performed a little worse than the Majority Baseline algorithm in correctly predicting whether a student was not attending class. In addition, the classifier performed a little better than the Majority Baseline algorithm in correctly predicting whether a student was attending class. Secondly, the Naïve Bayes algorithm performed equal to the Majority Baseline algorithm in correctly predicting whether a student was or was not attending class. Both algorithms were not able to correctly predict whether students were attending class. This was expected from the Majority Baseline algorithm, but not from the Naïve Bayes algorithm. Lastly, the Random Forest algorithm performed significantly worse than the Majority Baseline algorithm in correctly predicting whether a student was not attending class. However, the classifier performed significantly better than the Majority Baseline algorithm in correctly predicting whether a student was attending class. Here it is shown that the Random Forest algorithm uses the most sophisticated formula for predicting unseen data since it is able correctly predict class attendance based on Class Name, Class Start Time, and Class End Time.

In sum, this paragraph provided results of Experiment 1. According to Table 4.1.1, the Random Forest algorithm is the best approach when predicting class attendance based on only Class Information data. In addition, the Logistic Regression algorithm and the Naïve Bayes performed badly in correctly predicting whether a student was attending class where the Random Forest algorithm performed significantly better in correctly predicting whether a student was attending class. It was expected that Experiment 2 would perform better than Experiment 1 in predicting class attendance. This will be discussed in the following paragraph.

### 4.2 Results Experiment 2: How well can class attendance be predicted based on GPS Location data?

The previous paragraph provided the results of predicting class attendance based on Class Information data. This chapter will show the results of predicting class attendance based on GPS Location data. Both experiments focused on data obtained at the moment that class attendance is determined. Meaning that data was obtained between the start and end time of the class. The features selected for analyzing the classification performance of the models for predicting class attendance based on GPS Location data were: students' Latitude and students' Longitude. The target variable was Attendance. The accuracy scores and F1-scores of the second experiment are shown in Table 4.2.1.

| **Classifier** | Accuracy (training) | Accuracy (test) | F1-score (training) | F1-score (test) |
|---|---|---|---|---|
| Majority Baseline | 0.56 | 0.53 | 0.72 | 0.69 |
| Logistic Regression | 0.56 | 0.54 | 0.71 | 0.69 |
| Naïve Bayes | 0.57 | 0.56 | 0.65 | 0.64 |
| Random Forest | 0.99 | 0.60 | 0.99 | 0.60 |

Table 4.2.1: *Performance accuracy scores and F1-scores of predicting students' class attendance based on students' GPS Location data.*

Based on the performance scores per algorithm, Table 4.2.1 displays the Logistic Regression algorithm, Naïve Bayes algorithm, and Random Forest algorithm performing overall better than the Majority Baseline algorithm, results that support the expectations. The Logistic Regression algorithm and the Naïve Bayes algorithm performed equally well, a result that contradicts the expectations. Furthermore, the Random Forest algorithm performed better than the Logistic Regression algorithm and the Naïve Bayes algorithm, a result that supports the expectations. However, the Random Forest algorithm performed better on the training set than on the test set, which is an indication of overfitting. The Logistic Regression algorithm and the Naïve Bayes algorithm displayed a better F1-score than the accuracy score. Taking the accuracy score into account, the approach of Random Forest algorithm yields the best classification performance. Furthermore, the confusion matrices of Experiment 2 per algorithm will be provided.

| | Predicted not Attending | Predicted Attending |
|---|---|---|
| Majority Baseline | | |
| Actual not Attending | 316 | 0 |
| Actual Attending | 278 | 0 |
| Logistic Regression | | |
| Actual not Attending | 314 | 2 |
| Actual Attending | 270 | 8 |
| Naïve Bayes | | |
| Actual not Attending | 288 | 28 |
| Actual Attending | 232 | 46 |
| Random Forest | | |
| Actual not Attending | 200 | 116 |
| Actual Attending | 124 | 154 |

Table 4.2.2: *Test set Confusion Matrix values per algorithm based on students' GPS Location data.*

Table 4.2.2 displays the confusion matrices per algorithm for Experiment 2. A clear variation between the complexities of the models is shown in the table. Firstly, the Logistic Regression algorithm performed a little worse than the Majority Baseline algorithm in correctly predicting whether a student was not attending class. However, the classifier performed a little better than the Majority Baseline algorithm in correctly predicting whether a student was attending class. Secondly, the Naïve Bayes algorithm performed worse than the Majority Baseline algorithm in correctly predicting whether a student was not attending class. However, the classifier performed better than the Majority Baseline algorithm in correctly predicting whether a student was attending class. The Naïve Bayes algorithm is able to outperform the Majority Baseline in correctly predicting whether a student was attending class with different predictors. The algorithm was able to correctly predict 46 times whether a student was attending class, where it could not correctly predict whether a student was attending class based on Class Information data. Lastly, the Random Forest algorithm performed significantly worse than the Majority Baseline algorithm in correctly predicting whether a student was not attending class. In addition, the classifier performed significantly better than the Majority Baseline algorithm in correctly predicting whether a student was attending class. Overall, the accuracy scores and F1-scores of Table 4.1.1 showed better performances than the accuracy score and F1-score of Table 4.2.1. In addition, the Logistic Regression algorithm and the Random Forest algorithm performed better in correctly predicting when students were attending class in Experiment 1. However, the Naïve Bayes algorithm performed better in correctly predicting when students were attending class in Experiment 2. These are results that contradict the expectations.

In sum, this paragraph provided results of Experiment 2. According to Table 4.2.1 the Random Forest algorithm is the best approach for predicting class attendance based on Class Information data. In addition, according to Table 4.2.2 the Random Forest algorithm is the best approach for predicting class attendance based on GPS Location data. Furthermore, the Logistic Regression algorithm and the Naïve Bayes algorithm performed badly in correctly predicting whether a student was attending class, where the Random Forest algorithm performed significantly better in correctly predicting whether a student was attending class. However, the results of the Random Forest algorithm indicated overfitting. Meaning that conclusions should be made with caution. It was expected that Experiment 2 would perform better than Experiment 1. However, the Logistic Regression and the Random Forest algorithms performed better in predicting class attendance based on Class Information data. Furthermore, it was expected that Experiment 4 would perform better than Experiment 2. This will be discussed in paragraph 4.4.

### 4.3 Results Experiment 3: How well can class attendance be predicted based on GPS Location data and WiFi Location data obtained prior to the start of the class?

The previous paragraph provided the results of predicting class attendance based on GPS Location data. The first two paragraphs of this section both focused on data obtained at the moment class attendance was determined. However, this paragraph will show the results of predicting class attendance based on GPS Location and WiFi Location data obtained prior to the start of the class. The major difference between this paragraph and the previous two paragraphs is that WiFi Location data is added as predictor. This is possible because the data is obtained one hour, three hours, and six hours prior to the start of the class. The Logistic Regression, Naïve Bayes, and Random Forest classifiers were applied three times in this experiment. Once for data obtained one hour prior to the start of the class, once for data obtained three hours prior to the start of the class, and once for data obtained six hours prior to the start of the class. The features selected for analyzing the classification performance of the models were: students' Latitude, students' Longitude, and students' WiFi Location. The target variable was Attendance. The accuracy scores and F1-scores are shown in Table 4.3.1.

| Classifier | Accuracy (training) | Accuracy (test) | F1-score (training) | F1-score (test) |
|---|---|---|---|---|
| Majority Baseline -1 | 0.55 | 0.56 | 0.71 | 0.71 |
| Majority Baseline -3 | 0.56 | 0.52 | 0.72 | 0.69 |
| Majority Baseline -6 | 0.55 | 0.53 | 0.71 | 0.69 |
| Logistic Regression -1 | 0.63 | 0.70 | 0.63 | 0.67 |
| Logistic Regression -3 | 0.58 | 0.55 | 0.60 | 0.57 |
| Logistic Regression -6 | 0.57 | 0.57 | 0.58 | 0.58 |
| Naïve Bayes -1 | 0.65 | 0.67 | 0.65 | 0.67 |
| Naïve Bayes -3 | 0.58 | 0.61 | 0.59 | 0.63 |
| Naïve Bayes -6 | 0.60 | 0.59 | 0.60 | 0.60 |
| Random Forest -1 | 0.99 | 0.94 | 0.99 | 0.94 |
| Random Forest -3 | 0.99 | 0.90 | 0.99 | 0.90 |
| Random Forest -6 | 0.99 | 0.90 | 0.99 | 0.90 |

Table 4.3.1: *Performance accuracy scores and F1-scores of predicting students' class attendance based on students' GPS Location data and WiFi Location data obtained prior to start of the class.*

Table 4.3.1 displays the Logistic Regression algorithm, Naïve Bayes algorithm and Random Forest algorithm for Location data obtained one, three, and six hours prior to the start of the class. Firstly, the accuracy scores of the Logistic Regression algorithms showed a better performance than the Majority Baseline algorithms. Results that support the expectations. However, the F1-scores of the Logistic Regression algorithms showed a worse performance than the Majority Baseline algorithms. Results that contradict the expectations. Furthermore, the experiment containing data obtained one hour prior to the start of the class performed better than the experiments containing data obtained three and six hours prior to the start of the class. A result that supports the expectations. However, there was no significant difference between the experiment containing data obtained three hours prior to the start of the class, and the experiment containing data obtained six hours prior to the start of the class. A result that contradicts the expectations. Secondly, accuracy scores of the Naïve Bayes algorithms showed a better performance than the Majority Baseline algorithms. Results that support the expectations. However, the F1-scores of the Naïve Bayes algorithms showed a worse performance than the Majority Baseline algorithms. Results that contradict the expectations. The Naïve Bayes algorithms performed overall better than the Logistic Regression algorithms. Results that support the expectations. The experiments containing data obtained one hour prior to the start of the class performed better than the experiments containing data obtained three and six hours prior to the start of the class. A result that supports the expectations. In addition, the model containing data obtained three hours prior to the start of the class performed better than the model

containing data obtained six hours prior to the start of the class. A result that supports the expectations. Lastly, the Random Forest algorithms performed significantly better than the Majority Baseline algorithms, the Logistic Regression algorithms, and the Naïve Bayes algorithms. These are results that support the expectations. The experiment containing data obtained one hour prior to the start of the class performed better than the experiments containing data obtained three and six hours prior to the start of the class. A result that supports the expectations. However, there was no difference between the model containing data obtained three hours prior to the start of the class, and the model containing data obtained six hours prior to the start of the class. A result that contradicts the expectations. The approach of the Random Forest algorithm containing data obtained one hour prior to the start of the class yields the best classification performance. Furthermore, the confusion matrices of Experiment 3 per algorithm will be provided.

| | Predicted not Attending | Predicted Attending |
|---|---|---|
| Majority Baseline | | |
| Actual not Attending (-1) | 315 | 0 |
| Actual Attending (-1) | 252 | 0 |
| Actual not Attending (-3) | 290 | 0 |
| Actual Attending (-3) | 264 | 0 |
| Actual not Attending (-6) | 293 | 0 |
| Actual Attending (-6) | 261 | 0 |
| Logistic Regression | | |
| Actual not Attending (-1) | 249 | 66 |
| Actual Attending (-1) | 124 | 128 |
| Actual not Attending (-3) | 213 | 77 |
| Actual Attending (-3) | 172 | 92 |
| Actual not Attending (-6) | 199 | 94 |
| Actual Attending (-6) | 144 | 117 |
| Naïve Bayes | | |
| Actual not Attending (-1) | 185 | 130 |
| Actual Attending (-1) | 58 | 194 |
| Actual not Attending (-3) | 131 | 159 |
| Actual Attending (-3) | 54 | 210 |
| Actual not Attending (-6) | 128 | 165 |
| Actual Attending (-6) | 64 | 197 |
| Random Forest | | |
| Actual not Attending (-1) | 299 | 16 |
| Actual Attending (-1) | 19 | 233 |
| Actual not Attending (-3) | 269 | 21 |
| Actual Attending (-3) | 34 | 230 |
| Actual not Attending (-6) | 268 | 25 |
| Actual Attending (-6) | 30 | 231 |

Table 4.3.2: *Test set Confusion Matrix values per algorithm based on students' GPS Location data and WiFi Location data prior to start of the class.*

Table 4.3.2 displays the confusion matrices per algorithm for Experiment 3 containing data obtained one, three, and six hours prior to the start of the class. Firstly, the three Logistic Regression algorithms performed worse than the Majority Baseline algorithm in correctly predicting whether a student was not attending class. However, the three Logistic Regression algorithms performed significantly better than the Majority Baseline algorithm in correctly predicting whether a student was attending class. It was expected

that when the delta between the moment of prediction and the moment the class started increased, the accuracy of predicting class attendance decreased. Meaning that this would also result in the same decrease in the confusion matrices. It can be seen that this is the case for correctly predicting that students were not attending class. However, the model containing data obtained six hours prior to the start of the class performed better in correctly predicting whether students were attending class than the model containing data obtained three hours prior to the start of the class. Secondly, the Naïve Bayes algorithms performed significantly worse than the Majority Baseline algorithms in correctly predicting whether a student was not attending class. However, the Naïve Bayes algorithms performed significantly better than the Majority Baseline algorithm in correctly predicting whether a student was attending class. Table 4.3.2 showed that when the delta of the moment of the obtained data increased, the correct predictions of students not attending class decreased. However, the model containing data obtained three hours prior to the start of the class was the best performing model in correctly predicting whether students were attending class. This is a result that contradicts the expectations. The model containing data obtained six hours prior to the start of the class performed worse than the model containing data obtained one hour prior to the start of the class, a result that supports the expectations. Lastly, the Random Forest algorithms performed worse than the Majority Baseline algorithm in correctly predicting whether a student was not attending class. However, the algorithms performed significantly better than the Majority Baseline algorithm in correctly predicting whether a student was attending class. Table 4.3.2 showed that the models containing data obtained three hours and six hours prior to the start of the class performed equally well in correctly predicting whether students were not attending class. In addition, the model containing data obtained six hours prior to the start of the class performed better in correctly predicting class attendance than the model containing data obtained three hours prior to the start of the class. It was expected that Experiment 3 would perform better than Experiment 1 and Experiment 2. The accuracy scores of the Logistic Regression, Naïve Bayes, and Random Forest algorithms performed better in Experiment 3 than in Experiment 1 and Experiment 2. A result that supports the expectations. However, F1-scores of the Logistic Regression algorithm in Experiment 1 and Experiment 2 performed better than the Logistic Regression algorithms in Experiment 3. In addition, the F1-scores of the Naïve Bayes algorithm in Experiment 1 performed better than the Naïve Bayes algorithms in Experiment 3. However, the F1-scores of the Random Forest algorithms in Experiment 3 show better results than the Random Forest algorithms in Experiment 1 and Experiment 2.

In sum, this paragraph provided results of Experiment 3. Three experiments were conducted. Based on the accuracy scores Experiment 3 performed better than Experiment 1 and Experiment 2. However, based on the F1-scores there were variations in the best performing experiment. The Random Forest algorithm containing data obtained one hour prior to the start of the class is the best approach when

predicting class attendance based on GPS Location data and WiFi Location data obtained prior to the start of the class. Therefore, the GPS Location data and WiFi Location data obtained one hour prior to the start of the class will be used in Experiment 4.

### 4.4 Results Experiment 4: How well can class attendance be predicted based on GPS Location data, WiFi Location data, and Class Information data?

The previous paragraph provided the results of predicting class attendance based on GPS Location and WiFi Location data obtained prior to the start of the class. This chapter will show the results of predicting class attendance based on GPS Location and WiFi Location data obtained one hour prior to the start of the class, and Class Information data. This data will be used in the fourth experiment because the best performing experiment of Experiment 3 contained data obtained one hour prior to the start of the class. The features selected for analyzing the classification performance of the models for predicting class attendance based GPS Location data, WiFi Location data, and Class Information data were: Class Name, Class Start Time, Class End Time, Class Location, students' Latitude obtained one hour prior to start class, students' Longitude obtained one hour prior to start class, and students' WiFi Location data obtained one hour prior to start class. The target variable was Attendance. The accuracy scores and F1-scores are shown in Table 4.4.1.

| **Classifier** | Accuracy (training) | Accuracy (test) | F1-score (training) | F1-score (test) |
|---|---|---|---|---|
| Majority Baseline | 0.55 | 0.56 | 0.71 | 0.71 |
| Logistic Regression | 0.70 | 0.74 | 0.70 | 0.72 |
| Naïve Bayes | 0.66 | 0.67 | 0.67 | 0.67 |
| Random Forest | 0.99 | 0.99 | 0.99 | 0.99 |

Table 4.4.1: *Performance accuracy scores and F1-scores of predicting students' class attendance based on GPS Location data and WiFi Location data obtained one hour prior to the start of the class and Class Information data.*

Based on the performance scores per algorithm, Table 4.4.1 displays the Logistic Regression algorithm, Naïve Bayes algorithm, and Random Forest algorithm performing better than the Majority Baseline algorithm. This is a result that supports the expectations. The F1-score of the Naïve Bayes algorithm showed a worse performance than the F1-score of the Majority Baseline algorithm, a result that contradicts the expectations. The Logistic Regression algorithm showed a better performance than the Naïve Bayes algorithm, a result that contradicts the expectations. Furthermore, the Random Forest algorithm performed significantly better than the Logistic Regression algorithm and the Naïve Bayes algorithm, a result that supports the expectations. Compared to the other three experiments, the algorithms

of Experiment 4 are the best performing models. A result that supports the expectations. However, the Naïve Bayes algorithm performed equally well in Experiment 3 containing data obtained one hour prior to the start of the class, a result that contradicts the expectations. According to Table 4.4.1, the approach of Random Forest yields the best classification performance. Furthermore, the confusion matrices of Experiment 4 per algorithm will be provided.

| | **Predicted not Attending** | **Predicted Attending** |
|---|---|---|
| Majority Baseline | | |
| Actual not Attending | 315 | 0 |
| Actual Attending | 252 | 0 |
| Logistic Regression | | |
| Actual not Attending | 248 | 67 |
| Actual Attending | 89 | 163 |
| Naïve Bayes | | |
| Actual not Attending | 186 | 129 |
| Actual Attending | 59 | 193 |
| Random Forest | | |
| Actual not Attending | 309 | 6 |
| Actual Attending | 1 | 251 |

Table 4.4.2: *Confusion Matrix values per algorithm based on GPS Location and WiFi Location data obtained one hour prior to the start of the class, and Class Information data.*

Table 4.4.2 displays the confusion matrices per algorithm for Experiment 4. Firstly, the Logistic Regression algorithm performed worse than the Majority Baseline algorithm in correctly predicting whether a student was not attending class. However, the classifier performed significantly better than the Majority Baseline algorithm in correctly predicting whether a student was attending class. Secondly, the Naïve Bayes algorithm performed significantly worse than the Majority Baseline algorithm in correctly predicting whether a student was not attending class. However, the classifier performed significantly better than the Majority Baseline algorithm in correctly predicting whether a student was attending class. Lastly, the Random Forest algorithm performed a little worse in correctly predicting whether a student was not attending class. However, the classifier performed significantly better than the Majority Baseline algorithm in correctly predicting whether a student was attending class. Furthermore, the Random Forest algorithm performed significantly worse in falsely predicting whether a student was attending or not attending class. Meaning that the algorithm was able to predict class attendance in almost all cases correctly. It was expected that Experiment 4 would be perform better than the other three experiments.

Overall, Experiment 4 performed better than the other three experiments. However, the Naïve Bayes algorithm performed equally well in Experiment 3 and Experiment 4. On the contrary, the Random Forest algorithm of Experiment 4 is overall the best performing model and is able to predict class attendance with 99% accuracy.

In sum, this paragraph provided results of Experiment 4. According to Table 4.4.1, the Random Forest algorithm was the best approach when predicting class attendance based on GPS Location and WiFi Location data obtained one hour prior to the start of the class, and Class Information data.

Overall, this section provided results on the performance per model based on accuracy scores and F1-scores. Most of the proposed models performed better than the Majority Baseline algorithm. The Random Forest algorithm was overall the best performing model for predicting class attendance. The following section will provide conclusions on the proposed results. Furthermore, limitations of the research and directions for future research will be proposed.

# Section 5: General Discussion and Conclusions

In the previous sections, four experiments for predicting class attendance were proposed and conducted. The results of these experiments were obtained in Section 4. In this section the results will be evaluated and conclusions will be drawn from this information.

## 5.1 Answers to the Research Questions

In Section 1, the problem statement of this thesis was proposed as being the following: *"Can class attendance be predicted based on sensor data and education data".* This thesis was divided into the following four sub-questions to answer the problem statement:

1. *How well can class attendance be predicted based on Class Information data?*
2. *How well can class attendance be predicted based on GPS Location data?*
3. *How well can class attendance be predicted based on GPS Location data and WiFi Location data obtained prior to the start of the class?*
4. *How well can class attendance be predicted based on GPS Location data, WiFi Location data, and Class Information data?*

Four experiments for predicting class attendance were proposed based on these four questions. The remainder of this section will discuss the conclusions drawn from the results of the proposed experiments to answer the research questions.

### 5.1.1 How well can class attendance be predicted based on Class Information data?

In the theoretical section of this thesis it was acknowledged that class attendance is an important feature for predicting and explaining students' performance. In this thesis three features were identified as possible predictors for class attendance. In this paragraph the results for Experiment 1 are discussed. Experiment 1 only considered the feature Class Information as predictor for class attendance. This experiment ignored possible influence of GPS Location data and WiFi Location data.

It was expected that the experiment containing only Class Information data for predicting class attendance would be the least well performing experiment proposed in this thesis. This assumption was based on the different properties of the prediction features. With regards to the other features, Class Information is a more generic property of a class and does not contain information on the location of individuals. Since predicting class attendance is executed per individual it seems logical to use student-specific location data over generic Class Information data.

The general assumption with regards to the algorithms was that the most complex one is the best performing algorithm. The three algorithms used in this thesis were 1) the Logistic Regression algorithm,

2) the Naïve Bayes algorithm, and 3) the Random Forest algorithm, ordered based on complexity indicating that the Random Forest algorithm is the most complex model. These three algorithms were compared against the Majority Baseline algorithm to evaluate performance. On one hand it was desired that the algorithm outperforms the Majority Baseline algorithm. On the other hand it was desired that the algorithm predicts class attendance with the highest accuracy. Therefore, the multiple algorithms were evaluated and it can be concluded that the Random Forest algorithm was the best performing algorithm in this experiment.

Class attendance can be predicted based on Class Information data with a 68% accuracy score using the Random Forest algorithm. This scores indicates that it is outperforming the Majority Baseline algorithm which had an accuracy score of 53%, supporting the expectation. It also indicates that only Class Information data is not sufficient to entirely explain and predict class attendance since it is not even close to a 100% accuracy score. Depending on the purpose of predicting class attendance, these insights are useful. It gives a rough and generic indication which can be used, for instance, for building class schedules. In sum, predicting class attendance based on Class Information data is useful for long term purposes since Class Information data is available a long time prior to the class. The Random Forest algorithm has a positive contribution to predict class attendance in an early stage.

5.1.2 How well can class attendance be predicted based on GPS Location data?

As mentioned, three features were identified as possible predictors for class attendance. In this paragraph the results for Experiment 2 are discussed. Experiment 2 only considered the feature GPS Location as predictor for class attendance. This experiment ignored possible influence of Class Information data and WiFi Location data.

It was expected that the experiment containing only GPS Location data for predicting class attendance would perform better than the experiment containing only Class Information data. This assumption was based on the difference in properties of the prediction features. With regards to the other features, GPS Location is a student-specific property which can be measured at different moments in time. These properties raised the assumption to better suit prediction of individual events of class attendance compared to the properties of Class Information data.

In the second experiment again three algorithms 1) the Logistic Regression algorithm, 2) the Naïve Bayes algorithm, and 3) the Random Forest algorithm were build and compared against the Majority Baseline algorithm. The three algorithms were evaluated and the Random Forest algorithm was the best performing model in this experiment based on the accuracy score. However, the model performed significantly better on the training data than on the test data, which is an indication of overfitting caused by the complexity of the Random Forest algorithm. The second best performing model was the Naïve

Bayes algorithm which did not show signs overfitting and is therefore also an algorithm to take into consideration.

Class attendance can be predicted based on GPS Location data with a 60% accuracy score using the Random Forest algorithm. This scores indicates that it is outperforming the Majority Baseline algorithm which had an accuracy score of 53%, supporting the expectation. However, the Random Forest algorithm indicated signs of overfitting. Therefore, class attendance can be predicted based on GPS Location data with a 54% accuracy score using the Naïve Bayes algorithm. It also indicates that only GPS Location data is not sufficient to entirely explain and predict class attendance since it is not even close to a 100% accuracy score. Furthermore, with the knowledge gained from Experiment 1 it can be concluded that the hypothesis is not supported. In addition, the practical value of predicting class attendance based on GPS Location data is very low since Class Information data is more accurate and earlier available. In later experiments the added value of using GPS Location data as a predictor of class attendance in combination with other predictors is shown.

5.1.3 How well can class attendance be predicted based on GPS Location data and WiFi Location data obtained prior to the start of the class?

It was stated before that three features were identified as possible predictors for class attendance. In this paragraph the results for Experiment 3 are discussed. Experiment 3 considered GPS Location data and WiFi Location data as predictors for class attendance. This experiment ignored possible influence of Class Information data.

It was expected that the experiment containing GPS Location data and WiFi Location data to predict class attendance would perform better than Experiment 2 in which only GPS Location was used as predictor. However, in the previous paragraph it was concluded that the experiment containing only Class Information data performed better than the experiment containing only GPS Location data. This conclusion makes it also relevant to compare Experiment 3 to Experiment 1. The expectation related to this sub-question was based on the assumption that a model with two predictors performed better than a model with one predictor. Furthermore, it was expected that GPS Location data and WiFi Location data containing data obtained one hour prior to the start of the class would perform better than the models containing GPS Location data and WiFi Location data obtained three and six hours prior to the start of the class. In addition, it was expected that the model containing GPS Location data and WiFi Location data obtained three hours prior to the start of the class would perform better than the model containing GPS Location data and WiFi Location data obtained six hours prior to the start of the class. These expectations are raised by the general idea that the shorter the time frame between the event and prediction the more likely that the student is preparing for or traveling to school.

In the third experiment the three algorithms were again compared and evaluated. Overall, the Random Forest algorithm significantly outperformed the Majority Baseline, Logistic Regression, and Naïve Bayes algorithms. This is applicable for all three variants of Experiment 3: 1) GPS Location data and WiFi Location data obtained one hour prior to the start of the class, 2) GPS Location data and WiFi Location data obtained three hours prior to the start of the class, 3) GPS Location data and WiFi Location data obtained six hours prior to the start of the class. The algorithm containing data obtained one hour prior to the start of the class was, as expected, the best performing model. However there were signs of overfitting in Experiment 2, there were no indications of overfitting in this experiment. Furthermore, the algorithm containing data obtained three hours prior to the start of the class performed equal to the algorithm containing data obtained six hours prior to the start of the class. A remarkable performance of the model containing data obtained six hours prior to the start of the class which is interesting since it is not expected that students are preparing or traveling towards the classroom six hours prior to the start of the class.

Class attendance can be predicted based on GPS Location data and WiFi Location data with a 94% accuracy score using the Random Forest algorithm containing data obtained one hour prior to the start of the class. This scores indicates that it is outperforming the Majority Baseline algorithm which had an accuracy score of 56%, supporting the expectation. Comparing the accuracy scores from Experiment 2 and Experiment 3 explains the added value of using WiFi Location data in this model. These results are even more remarkable since GPS Location data and WiFi Location data obtained one hour prior to the start of the class is used instead of GPS Location data obtained on the moment the class is given. WiFi Location data adds value to the experiment by increasing the accuracy of class attendance prediction. This can be derived from the fact that WiFi Location data was used to determine class attendance and there is only a one hour difference between the prediction and the determination of class attendance. In sum, the increasing accuracy score shows that the model containing GPS Location data and WiFi Location data obtained one hour prior to the start of the class outperformed all previously discussed models. The practical limitation for this model is similar to Experiment 2 since data can only be obtained close to the class.

#### 5.1.4 How well can class attendance be predicted from GPS Location data, WiFi Location data, and Class Information data?

As determined in the previous paragraphs, three features were identified as possible predictors for class attendance. In this paragraph the results for Experiment 4 are discussed. Experiment 4 considered the features GPS Location and WiFi Location data obtained one hour prior to the start of the class, and Class Location data. This experiment ignored possible influence of GPS Location and WiFi Location data

obtained three and six hours prior to the start of the class. From here on, when mentioning GPS Location data and WiFi Location data, data obtained one hour prior to the start of the class is meant.

It was expected that the experiment containing GPS Location data, WiFi Location data, and Class Location data for predicting class attendance would perform better than all previously discussed experiments. The assumption that a model with three predictors will perform better than a model with two predictors is the base for this expectation, similar to the assumption in Experiment 2. Besides one additional predictor was used, also predictors with different properties were used. Students-specific and class-specific properties were combined in this experiment, raising the expectation that this experiment would outperform other experiments.

In the last experiment the best performing algorithm based on accuracy score was again the Random Forest algorithm, which supports the expectation that the most complex algorithm is also the most suitable algorithm for this thesis. However there were signs of overfitting in Experiment 2, there were no indications of overfitting in this experiment.

Class attendance can be predicted based on GPS Location data, WiFi Location data, and Class Location data with a 99% accuracy score using the Random Forest algorithm. This score indicates that it is outperforming the Majority Baseline algorithm which had an accuracy score of 56%, supporting the expectation. The reason for this high accuracy score is the introduction of the feature Class Location data which is used in combination with WiFi Location data to determine class attendance. The difference of determining class attendance and predicting class attendance in this experiment is only one hour. The limitation of this experiment is similar to the limitation of Experiment 3 with the exception that part of the prediction can already be executed when Class Information data is available, and the overall performance of the model which can only be executed one hour prior to the start of the class has also slightly increased compared to Experiment 3. In sum, the Random Forest algorithm using all three predictors is the best performing algorithm for predicting class attendance.

Concluding from these sub-questions, the main problem statement can now be answered. When answering the question "*Can class attendance be predicted based on sensor data and education data",* multiple aspects have to be taken into account. From the four experiments executed in this thesis the general conclusion can be drawn that class attendance can be predicted based on sensor data and education data. When interpreting this conclusion the accuracy of the prediction has to be taken into consideration. The accuracy score can differ per experiment and per algorithm which is shown in the Results Section. Overall, the Random Forest algorithm using all three predictors (i.e., GPS Location data, WiFi Location data, and Class Information data) is the best performing algorithm. Another aspect of the conclusion to take into account is the availability of data used for prediction. Since sensor data (i.e., GPS

Location data and WiFi Location data) is available in a later stage than education data (i.e, Class Information data) this will influence the practical use of this best performing algorithm. This improved model has a positive contribution in predicting class attendance at the moment Class Information data is available, improving the long-term class scheduling by Universities. In addition, for short-term purposes this model can predict class attendance with a high accuracy which enables professors to intervene earlier during a course and improve students' performances.

### 5.2 Directions for Further Research

The previous paragraph provided answers on the research questions to answer the problem statement and presented the conclusion based on the performances of the four proposed experiments. This paragraph will focus on discussing the limitations of this thesis and provides recommendations for future research.

Although the dataset is large and contains many features, the number of students participating in the study is small (N=48). "Such a small dataset is limiting because we cannot use more sophisticated predictive models or features because it may lead to overfitting", Wang et al. (2014). Furthermore, this dataset is not homogeneous because students mostly follow different courses. The students in the sample were not all computer science majors. However, all the students in the sample took one class they all had in common (i.e., Android programming). It is therefore difficult to generalize the results to other students at the Dartmouth University or to students in general. It would be interesting to investigate whether the proposed four experiments perform equally well when all courses are followed by the same 100 students.

Predicting class attendance based on 24 hours, 48 hours, 72 hours, or even one week before the start of the class would be interesting to analyze. The added value of being able to predict class attendance based on more than 24 hours prior to the start of the class offers new opportunities. For instance class schedule date/time changes or even class cancellations. In addition, it would be interesting to perform validation of the class schedule. The knowledge that the person is not attending class is obtained, however, it is not known where the student is. Furthermore, in the experiments GPS Location data was added as a predictor the training set performed very well with accuracy and F1-scores of 99%. At this moment, there is no clear explanation for this event. It would be interesting to investigate this more extensively to grasp the influence of GPS Location data on predicting class attendance.

Lastly, k-fold cross-validation was applied to validate the training set. However, cross-validation was not used to control for overfitting. The performance of the model was validated and the mean value, minimum value, and maximum value were calculated to address this. However, it would be more sufficient to split the data in training, validation, and test set. By doing this, it would be possible to control for overfitting and to control for data being trained on all examples. The major limitation of this thesis is that it is not controlled for the fact that input data can be only trained on a single class in the dataset.

Meaning that it is a possibility that the proposed models in this thesis are trained on one class which might lead to overfitting.

Overall, this study contributes within the existing framework by building further on the study by Zhou et al. (2016) who used WiFi Location data to analyze and determine class attendance. This study further explores that principle by predicting class attendance based on GPS Location data, WiFi Location data, and Class Information data. In addition, class attendance was predicted by using machine learning techniques and this thesis proposed three analyses on predicting class attendance based on different moments in time prior to the start of the class.

# References


Acharya, S. (2003). Factors affecting stress among Indian dental students. *Journal of dental education*, *67*(10), 1140-1148.

Anuradha, C., & Velmurugan, T. (2016, January). Feature Selection Techniques to Analyse Student Academic Performance using Naïve Bayes Classifier. In *The 3rd International Conference on Small & Medium Business* (pp. 345-350).

Attewell, P., Monaghan, D., & Kwong, D. (2015). Data mining for the social sciences: An introduction. *Univ of California Press.*

Bennedsen, J., & Caspersen, M. E. (2007). Failure rates in introductory programming. *AcM SIGcSE Bulletin*, *39*(2), 32-36.

Carroll, R. A., & Peter, C. C. S. (2017). Proportion of available points predicts student attendance in college courses. *The Psychological Record*, *67*(1), 61-69.

Chen, J., & Lin, T. F. (2008). Class attendance and exam performance: A randomized experiment. *The Journal of Economic Education*, *39*(3), 213-227.

Cilesiz, S. (2015). Undergraduate students' experiences with recorded lectures: towards a theory of acculturation. *Higher Education*, *69*(3), 471-493.

Conijn, R., Kleingeld, A., Matzat, U., Snijders, C., & van Zaanen, M. (2016). Influence of course characteristics, student characteristics, and behavior in learning management systems on student performance.

Cuseo, J. (2007). The empirical case against large class size: Adverse effects on the teaching, learning, and retention of first-year students. *The Journal of Faculty Development*, *21*(1), 5-21.

Davey, C., Aiken, A. M., Hayes, R. J., & Hargreaves, J. R. (2015). Re-analysis of health and educational impacts of a school-based deworming programme in western Kenya: a statistical replication of a cluster quasi-randomized stepped-wedge trial. *International journal of epidemiology*, *44*(5), 1581-1592.

Eagle, N., Pentland, A. S., & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, *106*(36), 15274-15278.

Friedman, P., Rodriguez, F., & McComb, J. (2001). Why students do and do not attend classes: Myths and realities. *College Teaching*, *49*(4), 124-133. doi: 10.1080/87567555.2001.10844593

Joshi, R. (2016). Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures. Retrieved from https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

Kassarnig, V., Bjerre-Nielsen, A., Mones, E., Lehmann, S., & Lassen, D. D. (2017). Class attendance, peer similarity, and academic performance in a large field study. *PloS one*, *12*(11), e0187078.

Le, J. (2018). A Tour of The Top 10 Algorithms for Machine Learning Newbies. Retrieved from https://towardsdatascience.com/a-tour-of-the-top-10-algorithms-for-machine-learning-newbies-dde4edffae11

Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and Predicting Students' Academic Performance Using Data Mining Techniques. *International Journal of Modern Education and Computer Science*, *8*(11), 36.

Mythili, M. S., & Shanavas, A. M. (2014). An Analysis of students' performance using classification algorithms. *IOSR Journal of Computer Engineering*, *16*(1), 63-9.

Nyamapfene, A. (2010). Does class attendance still matter? *Engineering education*, *5*(1), 64-74.

Statistics How To (2018). Z-Score: Definition, Formula and Calculation. Retrieved from http://www.statisticshowto.com/probability-and-statistics/z-score/#Whatisazscore

Steven E. Gump (2010) The Cost of Cutting Class: Attendance As A Predictor of Success, College Teaching, 53:1, 21-26, doi: 10.3200/CTCH.53.1.21-26

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., & Campbell, A. T. (2014, September). 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* (pp. 3-14). ACM.

Wang, R., Wang, W., daSilva, A., Huckins, J. F., Kelley, W. M., Heatherton, T. F., & Campbell, A. T. (2018). Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 43.

Westerman, J. W., Perez-Batres, L. A., Coffey, B. S., & Pouder, R. W. (2011). The relationship between undergraduate attendance and performance revisited: Alignment of student and instructor goals. *Decision Sciences Journal of Innovative Education*, *9*(1), 49-67.

Worthington, D. L., & Levasseur, D. G. (2015). To provide or not to provide course PowerPoint slides? The impact of instructor-provided slides upon student attendance and performance. *Computers & Education*, *85*, 14-22.

Zhou, M., Ma, M., Zhang, Y., SuiA, K., Pei, D., & Moscibroda, T. (2016). EDUM: classroom education measurements via large-scale WiFi networks. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (pp. 316-327). ACM.

# Appendix

| Student | Number of Classes per Student | Number of Attended Classes per Student |
|---|---|---|
| u01 | 98 | 1 |
| u02 | 68 | 2 |
| u03 | 56 | 1 |
| u04 | 31 | 0 |
| u05 | 99 | 2 |
| u07 | 70 | 8 |
| u08 | 97 | 22 |
| u09 | 35 | 1 |
| u10 | 96 | 17 |
| u12 | 106 | 2 |
| u13 | 59 | 18 |
| u14 | 82 | 20 |
| u15 | 64 | 14 |
| u16 | 60 | 15 |
| u17 | 100 | 20 |
| u18 | 125 | 8 |
| u19 | 105 | 15 |
| u20 | 36 | 2 |
| u22 | 87 | 1 |
| u23 | 33 | 1 |
| u24 | 54 | 2 |
| u25 | 51 | 9 |
| u27 | 107 | 8 |
| u30 | 90 | 14 |
| u31 | 83 | 2 |
| u32 | 75 | 33 |
| u33 | 65 | 0 |
| u34 | 30 | 8 |
| u35 | 58 | 7 |
| u36 | 0 | 0 |
| u39 | 0 | 0 |
| u41 | 55 | 1 |
| u42 | 39 | 13 |
| u43 | 26 | 0 |
| u44 | 66 | 3 |
| u45 | 25 | 7 |
| u46 | 91 | 2 |
| u47 | 35 | 2 |
| u49 | 57 | 2 |
| u50 | 38 | 0 |
| u51 | 64 | 22 |
| u52 | 95 | 12 |
| u53 | 67 | 14 |
| u54 | 60 | 3 |
| u56 | 0 | 0 |
| u57 | 71 | 6 |
| u58 | 60 | 4 |
| u59 | 117 | 7 |

Table A1*: Number of classes and number of attended classes per student*

| Student | Classes |
|---|---|
| u01 | ENGS 069, ENGS 022, ANTH 012 |
| u02 | COSC 077, COSC 098, COSC 065 |
| u03 | COSC 057, COSC 065 |
| u04 | COSC 065 |
| u05 | COSC 050, PSYC 028, COSC 065 |
| u07 | COSC 077, COSC 060, COSC 065 |
| u08 | CHIN 062, COSC 089 1, COSC 065 |
| u09 | ANTH 050, COSC 065, COSC 099 |
| u10 | COSC 050, BIOL 004, COSC 065 |
| u12 | COSC 089 1, COSC 050, TUCK 003, COSC 065 |
| u13 | COSC 070, COSC 065 |
| u14 | COSC 065, COSC 027, COSC 020 |
| u15 | EARS 003, SPAN 003, COSC 065 |
| u16 | COSC 065, COSC 027 |
| u17 | COSC 089 1, MUS 016, COSC 065 |
| u18 | COSC 089 1, CHIN 033, TUCK 003, COSC 065 |
| u19 | COSC 050, COSC 065, FILM 051 |
| u20 | COSC 070, COSC 065 |
| u22 | COSC 050, NAS 035, COSC 065 |
| u23 | COSC 070, COSC 065 |
| u24 | M&SS 045, COSC 060, ENGL 028 |
| u25 | NAS 008, COSC 065, ENGL 028 |
| u27 | ECON 024, JAPN 033, FILM 042, COSC 065 |
| u30 | MUS 003, COSC 050,  COSC 065 |
| u31 | MUS 003, COSC 077,  COSC 065 |
| u32 | MATH 023, COSC 069, COSC 065 |
| u33 | ENGS 025, ENG 069, ENGS 093 |
| u34 | COSC 070, COSC 065 |
| u35 | COSC 070, COSC 065 |
| u36 | |
| u39 | |
| u41 | COSC 050, SPAN 002, COSC 065 |
| u42 | COSC 070, COSC 065 |
| u43 | ENGS 031, COSC 065 |
| u44 | COSC 060, COSC 065 |
| u45 | COSC 070, COSC 065 |
| u46 | ECON 036, COSC 050, COSC 065 |
| u47 | COSC 060, COSC 065 |
| u49 | MATH 013, LAT 003, COSC 065 |
| u50 | COSC 060, COSC 065 |
| u51 | COSC 070, COSC 065 |
| u52 | BIOL 006, COSC 050, COSC 065 |
| u53 | COSC 089 1, COSC 065 |
| u54 | ENGS 025, MATH 023, COSC 065 |
| u56 | |
| u57 | COSC 098, COSC 060, COSC 065 |
| u58 | COSC 070, COSC 065 |
| u59 | GERM 001, COSC 065, COSC 007 |

Table A2*: Explanation of classes taken per student, provided by StudentLife.*

| Classes |
|---|
| ANTH 012 |
| BIOL 004 |
| BIOL 006 |
| CHIN 033 |
| CHIN 062 |
| COSC 007 |
| COSC 020 |
| COSC 027 |
| COSC 050 |
| COSC 057 |
| COSC 060 |
| COSC 065 |
| COSC 069 |
| COSC 070 |
| COSC 077 |
| COSC 089 1 |
| COSC 089 2 |
| COSC 089 3 |
| EARS 003 |
| ECON 024 |
| ECON 036 |
| ENGL 028 |
| ENGL 047 |
| ENGL 067 |
| ENGS 022 |
| ENGS 025 |
| ENGS 069 |
| ENGS 093 |
| FILM 051 |
| GERM 001 |
| HIST 051 |
| JAPN 033 |
| LAT 003 |
| M&SS 045 |
| MATH 013 |
| MATH 022 |
| MATH 023 |
| MUS 003 |
| MUS 016 |
| NAS 008 |
| NAS 035 |
| PSYC 028 |
| REL 018 |
| SPAN 003 |
| TUCK 003 |

Table A3*: Explanation of classes provided by Dartmouth University, according to the StudentLife dataset.*