# PREDICTING STUDENTS' GPA FROM MOBILE USAGE BEHAVIOUR AND WELL-BEING

MARIA NEDEVA

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

2100835

COMMITTEE

dr. Drew Hendrickson
dr. Lisanne Huis in't Veld

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 23th, 2023

WORD COUNT

7823

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor, Dr. Drew Hendrickson, for his guidance throughout this project. The provided expertise and support have been instrumental in the completion of this study.

# PREDICTING STUDENTS' GPA FROM MOBILE USAGE BEHAVIOUR AND WELL-BEING

MARIA NEDEVA

## Abstract

Education plays a significant role in an individual's life. It serves as a stepping stone, preparing students for future professional development. The need to study students' experiences during their academic journey has been recognized by various research disciplines. The results have been used to improve the educational process, identify and assist students at risk, create individual study plans, and uncover the social and personal factors that influence students' final grades. Extensive research has been conducted on applying machine learning algorithms to predict students' GPAs. One example is the field of Educational Data Mining, which utilizes predictive modeling to forecast students' grades based on academic data.

This research paper focuses on a similar approach, aiming to predict students' academic performance in terms of GPA (grade point average) by analyzing their mobile usage behavior and well-being. The primary objective of this study is to address the question: How accurately can mobile usage behavior and well-being predict students' grades? The distinguishing feature of this work is the utilization of the aforementioned predictors. By incorporating these factors into the prediction model, deeper insights can be gained into latent behaviors that affect academic achievements. The dataset used in this study consists of mobile phone usage logs, well-being surveys, and academic records. These data sources provide an overview of behavioral patterns, specific application usage, and mental well-being throughout the semester.

The main findings of this research demonstrate that mobile usage and well-being have an influence on students' academic performance. However, further analysis and the inclusion of more informative features will be necessary for future investigations. The following machine learning models were compared: Linear Regression, Support Vector Regression, Decision Tree, Random Forest, and Xgboost. The analysis shows that Random Forest most accurately captures the relationship between the target and the predictors, with a Mean

Absolute Error (MAE) of 0.631. However, additional model training and more robust optimization will be necessary to obtain results that are closer to the ground truth.frontmatter.tex

# 1 INTRODUCTION

The primary objective of this research is to investigate how accurately students' GPAs can be predicted from mobile usage behaviours and their well-being, by extracting features from raw phone data. Secondly, to identify the features that most influence the target variable. To achieve this, several machine learning models were tested. The best-performing one is selected based on Mean Absolute Error (MAE) score. This study aims to provide insight into the potential of utilizing mobile usage behaviour as a predictor of students' grades.

## 1.1 *Problem Statement*

Education plays a vital role in the personal and professional development of individuals. The exploration of students' experiences and academic performance has been recognized in the field of social sciences. Previous research has investigated various factors that influence students' academic performance, including social life, academic activities, socio-economic status, culture, parents' background, and environmental factors.

In recent years, the field of Educational Data Mining (EDM) in computer science has emerged, utilizing machine learning techniques to analyze and predict outcomes in students' academic performance, analysing educational data. However, alternative approaches have also examined the impact of variables such as well-being, mobile usage, and internet usage on students' academic achievements (Mukta et al., 2022; Wang et al., 2015; Xu et al., 2019). This broader range of variables reflects the increasing availability of diverse data sources, offering new ways for addressing predictive tasks related to human behaviours.

## 1.2 *Social and Scientific Relevance*

This study aims to contribute to the existing body of literature by exploring the relationship between mobile phone usage, well-being, and academic performance. Mobile phones have become pervasive in our daily lives, serving as tools that can both aid and distract us from various activities. Investigating their utilization by students and their influence can provide educational institutions with a better understanding of students' engagement and attention during lectures, as well as shed light on potential phone addictions.

From a scientific and analytical perspective, inferring informative features from raw mobile data presents new opportunities for analyzing such types of data. Moreover, this research offers the chance to validate and compare existing models in predicting phenomena based on human behavior infered from mobile usage.

Another application of the present research is to highlights the potential for creating a mobile app similar to the app SmartGPA used in Wang et al. (2015) or Human-in-the-Loop Cyber-Physical Systems (HiLCPS) that offers personalized learning methodologies, used in Sinche et al. (2020) study. Such an application could record students' behavior and well-being, utilizing the gathered data to estimate their potential grades. This innovative approach can be utilized by students for enhancing their academic performance and providing valuable insights into the factors that contribute to their success.

## 1.3 *Research Strategy*

The main research question defined below bears its motivation from the existing research gap in the literature and not enough data from scientific research on mobile phone usage and well-being as a predictor of academic performance.

> *How well students' academic grades can be predicted from their daily behaviors of phone usage, sleep and study habits, university visits, and reported well-being using machine learning models, and how different models capture the relationships in the data?*

The following sub-research questions serve as a guideline for answering the above-stated problem:

RQ1 *Which machine learning model can most accurately capture the relationship between the target variable and the predictors?*

- For the purpose of this question the following machine learning algorithms were employed and compared: Linear Regression, Support Vector Regression, Decision Tree, Random Forest and Extreme Gradient Boost (Xgboost).
- The models were evaluated in the cross-validation stage and on the test set with Mean Absolute Error.

RQ2 *What predictors identified by Shapley values contribute to the best-performing model's predictions?*

RQ3 *What are the patterns of error across all the models in predicting students' grades?*

## 2 RELATED WORK

The literature review is divided into two sections: 1) A brief introduction to educational data mining, and 2) An academic performance analysis with non-educational parameters. The two sections account for the methods used, datasets and results obtained.

### 2.1 *A brief introduction to educational data mining*

Educational data mining (EDM) is a relatively new field in the information sciences. As the name suggests, the data created in the educational institution's context is the main information source. Additional focus on the socio-economic background can also be found in most of the studies(Abu Saa et al., 2019). Predicting the students' performance (in terms of grades) is not a new topic to EDM. Various types of research have been carried out in classifying or regressing the final grade. Batool et al. (2023) performs comparative analysis to explore the various techniques for analyzing academic data. He points out that the most used models in the field of EDM are Decision Tree, Naive Bayes, Neural Networks, Support Vector Machines, K-Nearest Neighbours, and Random Forest. Another similar review on predictive modelling for EDM conducted by Zhang et al. suggested that the most used methods for regression analysis are Linear Regression, Support Vector Regression, Deep Learning, and Markov Network. Additionally, models such as decision trees and linear regression are more promising due to the ease of interpretation.

In one study Yağcı used midterm grades, department data and faculty data to predict the final grade. They employed Random Forest, Artificial Neural Network, Logistic Regression, Support Vector Machines, Naive Bayes, and K-Nearest Neighbours, where Random Forest and Artificial Neural Network show the highest accuracy of 74%. In another study Gadhavi and Patel applied Linear Regression on historic grades to predict the final grade. Abu Saa et al. (2019) analyze students' demographic backgrounds and social life using Decision Trees and Naive Bayes. They found that students' academic performance can be influenced by other factors that are not related to their educational activities. A similar approach has been carried out with predicting the final grade(El Aissaoui et al., 2020). The authors use Multiple Linear Regression on demographic data. Gaftandzhieva et al. (2022) predicts final grades based on activities on an e-learning platform and attendance in online lectures. They train several models: Random Forest, Extreme Gradient Boosting, K-Nearest Neighbours and Support Vector Machines. In their analysis, Random Forest performance is with the highest accuracy of 78%.

## 2.2 *Academic performance analysis with non-educational parameters*

A few alternative approaches investigate the relationship between well-being and academic performance. Mukta et al. (2022) used MPNet to estimate students' psychological attributes and mental well-being from their Facebook newsfeed. The inferred new variables are used for predicting academic performance and an accuracy of 94% is reported. In their research, the student performance is divided into two groups: "students with good performance" and "students with poor performance", hence, students were classified into these two groups. The findings showed that well-being is correlated to academic performance and also can be used as a predictor of the second. Two studies support the findings of a correlation between academic performance and experiencing different mind states(in the form of beliefs, emotions and well-being)(Kotzé & Kleynhans, 2013; Mubarok & Pierewan, n.d.). Kotzé and Kleynhans (2013) find a significant correlation between burnout, emotional exhaustion and cynicism and academic performance after conducting Pearson product-moment correlation analysis. Mubarok and Pierewan (n.d.) examined the importance of well-being as a predictor, with the use of regression analysis. The study was conducted on adolescents in high school. The main findings suggest that students who have good well-being also have good academic achievements. However, no further studies on training machine learning models with well-being data have been performed for estimating academic grades.

Xu et al. (2019) trained different machine learning models to predict students' performance based on internet usage data (online time, offline time, download volume, upload volume, terminal device)(Xu et al., 2019). They estimated the academic performance of 72% accuracy with Support Vector Machines. The findings revealed that students' grades can be predicted by patterns in internet usage. Yet, no further investigation has been carried out on the content consumed and how the internet was utilized. Rajalaxmi et al. (2019) tried to predict the grade of students in engineering disciplines. The authors infer several predictors related to internet usage: usage of the internet for educational purposes, usage of the internet for entertainment purposes, utilization of the internet for communication purposes, active duration in social media networks, and usage of the internet before the final exams (Rajalaxmi et al., 2019). Multivariate Linear Regression was applied to subsamples of the data where each subsample represented a particular engineering discipline. Furthermore, they explore the result across different grade ranges: RMSE of 0.24 for grades between 7.1 and 8.0, RMSE of 0.3 for grades in the range 8.1 - 9.0, and RMSE of 0.13 for grades between 9.1 - 10.0.

The SmartGPA study from 2015 conducted by Wang et al. explored the influence of a wide range of behaviours on students' academic performance. The data was collected through a mobile application that records passive sensor data such as phone activity, location, accelerometer, sleep duration, quality of sleep, and etc. (Wang et al., 2015). Likewise, students were asked to report their well-being and daily mood. Additional features were engineered from the raw data. One example is study behaviour. The authors used these data to identify locations where students spent a significant amount of time, such as libraries, dorms, and classrooms. They also exploit the data to identify periods of time when students were stationary for more than 20 minutes, which they defined as studying. Similarly from the conversation recorder and accelerometer, social life was measured. Furthermore, behavioural change was captured by using behavioural slope and behavioural breakpoints. The behavioural slope was applied to capture the magnitude of change (increase or decrease in sleep). Behavioural breakpoints captured the specific time when change occurs (Wang et al., 2015). Linear Regression with Lasso Regularization, Support Vector Regression and Decision Trees were employed to the data. However, Linear regression outperformed the other two models. Due to the limited amount of data (N=30), the more complex models did not perform well (Wang et al., 2015). A similar approach was used in another more recent study. Sinche et al. (2020) analyzed student performance using Human-in-the-Loop Cyber-Physical Systems (HiLCPS) that create personalized study recommendations. They explored the correlation between academic performance and inferred behaviour from mobile sensor data. Similarly to the above-mentioned study, additional variables were inferred such as the average time students spend at university, sociability measured from the amount of incoming and outgoing calls, physical activity, etc. The study suggests that academic performance can be influenced by these factors. However, no machine learning predictions have been performed. The applications of passive sensor data in combination with self-reported data have been used in the field of mental health studies (Bai et al., 2021; Jacobson & Chung, 2020; Pratap et al., 2019). Results show that daily behaviours related to phone usage and mobility can be used for classifying patients with depressive disorders. Nevertheless, features extracted from GPS data in combination with app event data and dwell time can give information about behaviours in particular places. Bai et al. (2021) found a correlation between phone usage routine, mobility, and depression among patients with major depressive disorder. They created daily windows with a range of 6 hours to measure the type of phone application used and dwell time. Several models were tested, but random forest shows the highest accuracy.

The lack of recent research in educational studies on predicting students' grades with extracted features from mobile sensor data and well-being reports motivates the approach for this paper. Moreover, the SmartGPA study from 2015 does not include mobile usage routine in their analysis, nor type of mobile activity, which according to the social sciences literature can be also correlated to academic performance (Giunchiglia et al., 2018; Hawi & Samaha, 2016). Giunchiglia et al. (2018) shows a negative impact of social media usage on academic performance. Moreover, phone inactivity during lectures influences the overall performance of students (Giunchiglia et al., 2018).

## 3 METHODOLOGY AND EXPERIMENTAL SETUP

To address the research problem of this study, the sub-sections below describe all the steps employed for data pre-processing, feature engineering, choice of models, evaluation and features contribution. Figure 1 below illustrates the data science research pipeline. New variables were engineered from the raw app events dataset, well-being dataset, one-time daily dataset. As previously mentioned, the SmartGPA study from 2015 observed students' various behaviours but did not incorporate mobile phone behaviours or types of activities performed on mobile phones as predictors. Moreover, Xu et al. (2019) examined internet consumption among students, but did not go into detail on the type of activities performed. Giunchiglia et al. (2018) 2018 and Hawi and Samaha (2016) have explored the relationship between social media app usage and academic performance, but no machine-learning predictions have been performed. Their findings motivate the decision for engineering additional variables related to monthly social media platform usage, instant messaging applications, streaming services, and internet browsing. Figure 21 in the Appendix A shows a bar plot of the most used application categories. The data shows that primary positions have the above-mentioned app categories. These results raised the hypothetical question: do the mainly used applications in these categories significantly affect the final grade? Section 3.3 introduces the constructed variables.

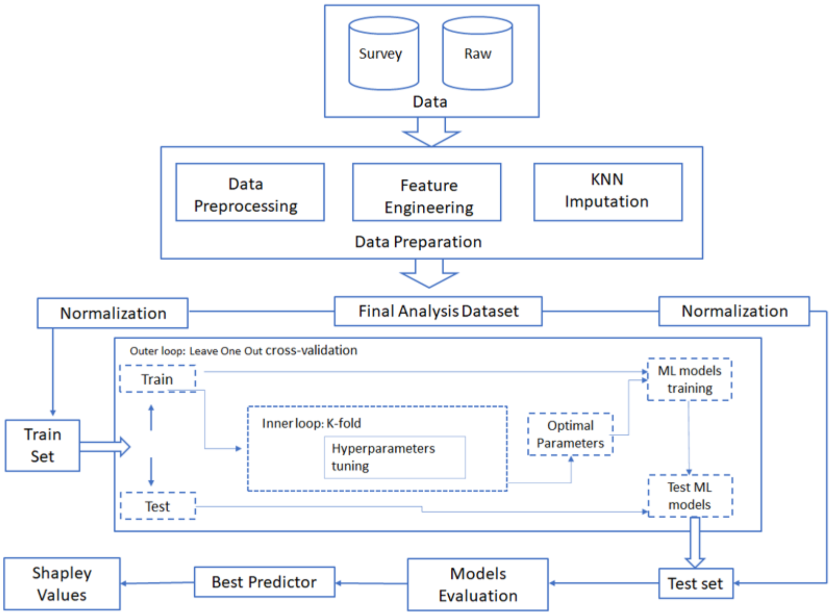


Figure 1: Data Science Research Pipeline

## 3.1 *Dataset Description*

The dataset used in this study is based on a prior investigation conducted among students at Tilburg University. The data collection period spans a duration of six months, commencing on January 23, 2020, and concluding on June 20, 2020. The data encompasses two primary categories of datasets: survey and mobile sensor data. The following paragraphs provide a more detailed description of the files.

### 3.1.1 *The one-time survey*

This table contains a comprehensive set of 136 questions, encompassing students' demographic attributes (age, sex, gender), technical details (phone model, data version), as well as a diverse aspect of students' psychological attributes related to Body Dysmorphic Disorder, Burnout, Major Depressive Disorder, Fatigue, Perceived Stress Scale, Procrastination, Connectedness, Big Five Personality traits, Impulsivity, Morningness/Eveningness, and Social Desirability. Each question within this table is assessed using a 5-point Likert scale, capturing the students' subjective responses to the given psychological constructs. The one-time survey was provided to the students only once for the whole period of the data collection.

### 3.1.2 *Once-daily survey*

The dataset comprises participants' records, capturing their daily responses to aspects of their sleep routine, including wake-up time, bedtime, sleep quality, and the duration it takes to fall asleep or wake up. Furthermore, the dataset includes a question regarding the daily time spent on studying activities.

### 3.1.3 *Five-times daily*

The five-times daily table is designed to assess the well-being of students. It consists of a set of questions covering topics such as stress, fatigue, procrastination, and happiness. In total, there are 10 questions included in this table. The subjective responses of the students were recorded using a 7-point Likert scale.

### 3.1.4 *Grades folder*

This folder contains individual files for each separate student. Each file represents a table that contains information about the number of courses taken by the respective student, along with the corresponding course names and the grades received for each course.

### 3.1.5 *App events folder*

The app event folder consists of individual files for each student. These files contain tables that include various attributes such as student ID, session, start time, end time, timestamps for start and end times, notifications, application details, battery usage, survey ID, longitude, and latitude. Additionally, columns with more descriptive application names and categories were incorporated from a supplementary table that provides category groups. The data is in a time series format, capturing observations over a period of 6 months, throughout a mobile application 'MobileDNA'. The total number of records amounts to 5,073,192. The recorded duration of the application log spans from 01-23-2020 to 06-20-2020.

The number of participants who completed all the surveys is N=236. An additional folder contains app event data from 10 students who did not answer any of the surveys. However, some of these students have an informative amount of app logs recorded. Appendix A contains EDA plots related to phone usage and well-being.

## 3.2 *Data Preprocessing*

### 3.2.1 *Data Cleaning and Data Imputation*

Initially, the tables from the App events folder were consolidated into a single file. This resulted in a total of 5,073,192 observations. However, certain students were excluded from the analysis based on specific criteria. Participants who had less than 200 sessions per month and who participated for less than 3 months. The decision was made on the fact that some features were aggregated on a monthly basis. Including participants with months less than the threshold would result in creating more missing values. After removing the participants, the total number of observations was estimated to be 4,632,923. That decreased the number of participants to N=193.

In addition, the location features in the dataset were found to contain corrupted values, requiring further detection and removal. To address this issue, the K-nearest neighbours (KNN) imputation was applied to impute the missing values created after the deletion. This technique utilizes the values of neighbouring data points to estimate and fill in the missing values in the dataset. More on the missing data imputation with KNN can be found in section 3.4.

The well-being dataset included data for months that exceeded the scope of the study. Figure 15 in the Appendix A illustrates the distribution of these months. It was hypothesized that this discrepancy was due to a system error. To ensure that important information was not lost, the values corresponding to the months outside the study scope were added to the nearest month in order to preserve the overall continuity of the data. The dataset contained missing values due to non-responses from some students on all the daily surveys. The missing values were around 30% and deletion was not a preferable technique. Again, KNN imputation was employed to deal with the missing data.

The one-time daily dataset also included 27% of missing data that was also imputed by utilizing KNN imputation.

### 3.2.2 *Categorical Encoding*

The one-time dataset contained a table of participants' gender. Binary encoding was used to transform the categories. 'Females' were set to 0, and 'Men' were set to 1. Figure 16 in the Appendix A illustrates the distribution of gender among students.

### 3.2.3 *Target variable transformation*

The lists of grades for each student were taken and the average grade point (GPA) was calculated. Figure 2 presents the distribution of the average grade among students.

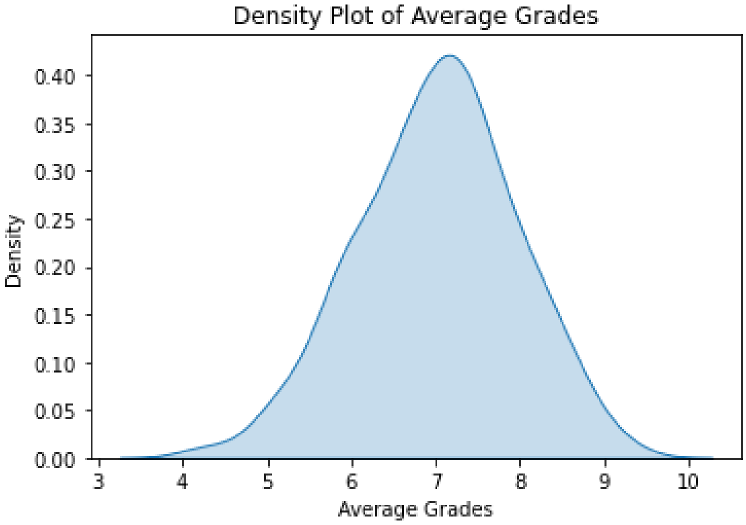


Figure 2: Distribution of the average grades

## 3.3 *Feature Engineering*

As was mentioned at the beginning of the Methods section, the raw app logs data required further feature engineering for inferring valuable information and to make it suitable for the models chosen. The raw locations were used to for mapping out the campus area. Additionally, five-times daily dataset and a one-time daily dataset were used for creating new variables. The next subsections describe the features engineered and the techniques employed.

### 3.3.1 *Trend slopes for measuring behavioural change*

Trend slopes were engineered by fitting a regression line through the data points for the selected variables. Table 1 provides an overview of the features for which trend slopes were computed. This approach was utilized in the SmartGPA study for measuring the correlation between behavioural change and final grade. The coefficient of the slopes shows a direction and strength in behavioural change over time. A positive value indicates an increase in change, on the other hand, a negative value reports a decrease in behaviour (Wang et al., 2015).

In general, three trend slopes were computed: 1) overall slope showing the change of behaviour from the beginning to the end of the study; 2) trend slope from the beginning of the study to the mid-semester (January – March); 3) and slope from the second part of the semester to the end of the study (April – June).

By utilizing these trend slopes, the study aimed to capture the influence of fluctuations in phone usage, well-being, study habits, and sleep patterns on the final grades achieved by the participants. The decision to split the data into two periods, pre and post-mid-semester, was based on the recognition that March and June encompass critical examination periods.

Table 1: Trend Slopes Features

| **Dataset** | **Features (mid-semester slope, post-mid-semester slope, overall slope)** |
|---|---|
| App events | Social Networking<br>Instant Messaging<br>Streaming Services<br>Internet Browsing<br>Email<br>Dialer<br>Overall phone usage |
| Well-being | Relaxed<br>Rushed<br>Stressed<br>Energy<br>Proactiveness<br>Concentration<br>Delay<br>Time waste<br>Procrastination<br>Happiness |
| Once-daily | Sleep hours<br>Study hours |

## 3.4 *Frequency of university visits and time spent on campus*

The longitude and latitude variables from the App event dataset were used to estimate the frequency of university attendance per month and the total time in minutes spent there. The initial feature engineering involved obtaining and filtering only the locations that fell within the campus area. The process included the following steps: 1) Campus geographic mapping: Google Maps was employed to delineate the geographical area of the campus. 2) Polygon representation: The Shapely library's Polygon feature was utilized to represent the campus space in the computational environment. 3) Location filtering: The dataset locations were iterated through the polygon representing the campus area, and a new boolean column was created to indicate whether a location falls within the campus scope

For the final analysis, two features were created. The number of university visits per month for each student was determined by summing the unique days per month where the boolean column indicated that the location is on campus. The total time in university per month was computed from the start time and end time stamps in the dataset.

### 3.4.1 *Ratio of app usage*

To evaluate students' engagement with specific phone applications, each day was partitioned into four distinct time windows: 6 am to 12 pm, 12 pm to 6 pm, 6 pm to midnight, and midnight to 6 am. This division was employed to assess students' interactions with the selected phone apps investigated by Giunchiglia et al. (2018) in their paper, as well as two additional apps, Dialer and Email, which were found to be extensively used by the students.

The aim behind partitioning the day into daily windows is to capture and measure students' levels of engagement with the identified apps during different periods of the day. The ratio of interactions with the mobile apps was calculated to measure the levels of engagement. Table 2 presents the categories used for computing the ratio.

Table 2: Ratio Features

| **Dataset** | **Features** |
|---|---|
| App events | Daily windows: 6 am to 12 pm, 12 pm to 6 pm, 6 pm to midnight, and midnight to 6 am |
| | Social Networking<br>Instant Messaging<br>Streaming Services<br>Internet Browsing<br>Email<br>Dialer<br>Overall phone usage |

## 3.5 *Final dataset consolidation and KNN imputation*

After combining all the tables, the subsequent statistical analysis revealed that more than 30% of the data was missing. Given the substantial percentage of missing information, the deletion of incomplete cases was considered inappropriate because can create loss of valuable information. Univariate methods, such as mean imputation, were considered as a potential alternative. However, imputing large proportions of missing data using methods like fill forward and back forward median imputation could introduce bias (Lodder et al., 2013).

To overcome that problem a more suitable algorithmic approach was required (Lodder et al., 2013). Therefore, the K-nearest neighbours (KNN) algorithm was employed to impute the missing data. KNN imputation utilizes the values of neighbouring data points to estimate the missing values based on their similarity in the feature space. By leveraging this algorithmic approach, the missing data was effectively imputed, mitigating the impact of missingness on subsequent analyses.

The final data set consisted of 243 variables and 193 observations The data was split on train and test set. After the split the training set was with 144 observations, while the test set was with 49 observations.

### 3.6 *Normalization*

The next necessary step before building the models was to position the data on the same scale. Normalization is a highly required step when the values in the data vary across scales, and therefore are placed in an extremely large space. Not normalizing or standardizing the data can lead to poor model performance and difficulties in capturing the relationships in the data. This process was performed using MixMaxScaler from the Sklearn library.

### 3.7 *Algorithms*

Several supervised machine learning algorithms were employed on the final dataset. The choice of models is motivated by their frequent use in the literature, both for predicting continuous outcomes and in educational studies for grade prediction. The following set of models was utilized: Mean prediction as a baseline model. Linear Regression (LR), Support Vector Regression (SVR), and Decision Tree (DT), these three models were used in Wang et al. (2015) and Xu et al. (2019) studies. The application of Random Forest (RF) and Extreme Gradient Boost (Xgboost) was motivated by the imperative to see how well more complex models perform on the type of data used for this study.

### 3.7.1 *Linear Regression*

Linear Regression is a commonly used statistical model in the field of Data Science, primarily for predicting continuous outcomes. The objective of the model is to find a relationship between the target and one or more independent variables, assuming a linear relationship. The aim of Linear Regression is to determine the line of the best fit that can capture the pattern in the data points. This process involves minimizing the difference between the actual values and the predicted values. This model was employed in Wang et al. (2015). In their work, additionally, Lasso Regularization was applied to remove redundant variables. This method outperforms the Decision Tree and Support Vector Regression. However, the volume of data used in the study is N=30, which makes Linear Regression a suitable choice, due to the low number of observations. The simplicity of the method makes it suitable for smaller datasets. The choice of this model is partly motivated by the above-mentioned study and the amount of the present data.

### 3.7.2 *Support Vector Regression*

Support Vector Machines (SVM) is a popular algorithm for both classification and regression. Support Vector Regression is the conversion of SVM for predicting continuous outcomes. The goal of the model is to find a function that approximates the relationship between the predictors and the predicted variable while minimizing the prediction error (Williamson et al., 2001). The process behind the model includes positioning the data points on a high-dimensional feature space between decision boundaries and inserting a hyperplane that has to minimize the prediction error. In comparison to linear regression, SVR exhibits robustness to outliers, with the help of the decision boundary that cuts off unusual values. This method shows the best performance in Xu et al. (2019) study. Part of the motivation behind using this model comes from the high-dimensional feature space, which makes it possible for the model to handle datasets with a larger amount of predictors. Moreover, SVR is suitable for a small number of observations. Table 4 in the Appendix B shows the hyperparameters used for optimization.

### 3.7.3 *Decision Tree*

The Decision Tree model is a recursive method that can be employed for both classification and regression tasks. The algorithm operates by constructing a series of rules, which are then used to partition the data into smaller subsets, resulting in a tree-like structure. Decision trees are suitable for capturing complex relationships in the data. A further advantage of the method is the ease to work with mixed types of data (continuous and categorical) without the need for additional pre-processing. Decision Trees can handle data that has larger dimensions by selecting the most informative features. Table 5 in Appendix B reports the set of hyperparameters used for the model's optimization.

### 3.7.4 *Random Forest*

Random forest is an ensemble model for both regression and classification. This model combines the predictions of multiple Decision Trees and based on their predictions provide outcomes. By randomly selecting a subset of features at each tree in the forest, Random Forest focuses on the most informative features, reducing the impact of noise and irrelevant variables. Once all the decision trees are built, Random Forest combines the predictions of each tree to produce a final estimation. This is done by averaging the predictions of all the trees. The model has shown superiority in a few of the studies Bai et al. (2021), Gaftandzhieva et al. (2022), and Gadhavi and Patel (2017) studies. Table 6 in Appendix B presents the set of hyperparameters used in the cross-validation stage.

### 3.7.5 *XGBoost*

Another ensemble model used in this study is XGBoost. Similarly to Random Forest, the model uses multiple decision trees. Conversely, the process behind utilizing the trees is different. The trees are built on top of each other to minimize the error from the previous tree. The basic principle is building a tree where in each node the similarity score is calculated. In other words, the first tree in the sequence tries to predict the output variable, while the subsequent trees in the sequence try to predict the residual errors left by the previous trees. The predictions of each decision tree in the sequence are then combined, using a weighted sum, to produce the final prediction. In comparison to Random Forest, XGBoost is considered as a more advanced algorithm that adds different weights to the tree leaves based on their contribution to the model, while RF assigns the same weights to all of the leaves. The set of hyperparameters used for model optimization can be found in Table 7 in Appendix B.

## 3.8 *Cross-validation and hyperparameter tuning*

Nested cross-validation is a method used for model building, assessment and selection. With two loops of cross-validation, the model is tuned in the inner loop, while its performance is evaluated in the outer loop. Figure 3 illustrates the process applied on the training set. For the specifics of the present data, Leave One Out Cross-Validation (Loocv) was utilized for the outer loop. And K-Fold Cross-Validation (KFoldcv) was used for Grid Search and hyperparameter tuning. This approach is used to solve the problem with bias and error by using the outer loop as unseen data that is not being included for the model optimization in the inner loop (Wainer & Cawley, 2021).

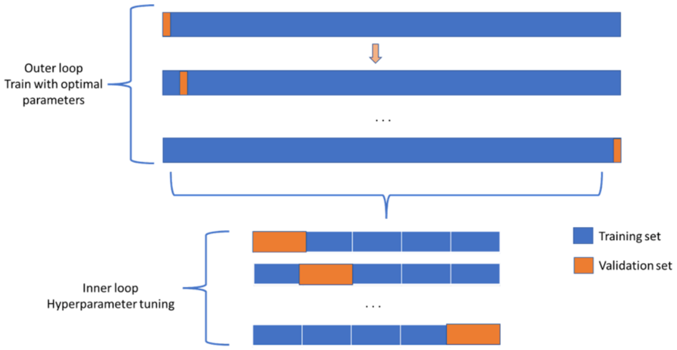


Figure 3: The process of nested cross-validation applied on the training set. The inner loop serves fro hyperparameter tuning, while the outer loop is used for model evaluation

### 3.9 *Evaluation*

Mean absolute error (MAE) was used to assess the models' performance on the cross-validation stage and on the hold-out set. MAE is a preferred choice over mean squared error (MSE) and root mean squared error (RMSE) when predicting students' grades. The metric is not affected by extreme values or outliers, making it a robust metric. It enables the estimation of absolute deviations from the ground truth and facilitates the selection of the model with the lowest average error.

R-squared was employed in the error analysis to evaluate how well the models fit the new data and explain it's variability. The scores from this metric were utilized to explain the error plots in the Results section.

### 3.10 *Predictors' contribution*

To determine the predictors that most affect the model Shapley values were computed. Shapley is an explainable machine learning technique that allows local and global analysis on the data and research problem. The local explainability suggests how each feature affects the result of every instance. On the contrary, the global importance is measured by averaging the absolute Shapley values of the feature contribution to the overall prediction. SHAP foundations come from the 'game theory' paradigm, and it is often described as a cooperative game where each feature is a player and the individual contribution of every player to the game output has been measured (Antwarg et al., 2019). Each feature contribution is calculated by: 1) Performing the original model prediction; 2) Removing the feature whose contribution we want to measure from the dataset and do a prediction without that feature; 3) The two predictions are being compared. The difference between the predictions describes the feature contribution for a particular instance; 4) Shapley value is calculated by averaging feature contributions over all possible feature subsets.

### 3.11 *Programming language and software*

The data processing was executed on Visual Studio Code using Python (Van Rossum & Drake Jr, 1995). The choice of programming language was made based on the variety of open-source libraries that are appropriate for data analysis and machine learning models. Pandas (pandas development team, 2020) library was used for the tables construction, and feature engineering in combination with Numpy (Harris et al., 2020) and other statistical modules. Seaborn (Waskom, 2021) and Matplotlib (Hunter, 2007) were used for the graph plotting. Scikit library (Pedregosa et al., 2011) was utilized for employing most of the machine learning models and grid search for hyperparameter tuning. Only Extreme Gradient Boosting model was imported from Xgboost library (Chai & Draxler, 2014). SHAP library (Lundberg et al., 2020) was used for Shapley values to understand features contribution. The Shapely library(Gillies et al., 2007–) was utilized for mapping out the campus area on a 2D space.

## 4 RESULTS

The results section covers subsections related to the models performance, comparison to the baseline model, and feature contribution to the model found to have the highest mean absolute error. Table 3 presents the results from the models' evaluation on the test set and the baseline predictions. The baseline simply calculated the mean grade. Random Forest is found to be the best-performing model in the case of this study, followed by Support Vector Regression with a small difference in the evaluation metrics coefficients. The next model ranked by it's performance is Decision Tree. These three models succeeded in outperforming the baseline model. The other two models, Linear Regression and XGBoost scored a higher value for MAE and a negative value for R-squared. Additional Table 8 with models' optimal hyperparameters can be found in the Appelndix B.

Table 3: Model Performance

| **Models** | **MAE** | **R-squared** |
|---|---|---|
| **Random Forest** | **0.631** | **0.226** |
| SVR | 0.639 | 0.221 |
| Decision Tree | 0.651 | 0.164 |
| XGBoost | 0.75 | -0.134 |
| Linear Regression | 1.143 | -1.386 |
| Baseline | 0.724 | -0.00225 |

## 4.1 *Models' performance evaluation and error analysis*

The following subsections covers the results from models' evaluation in terms of MAE and R-squared. Additionally, plots illustrating the results from the error analysis for each model are present.

### 4.1.1 *Random Forest performance*

Table 3 presents models' results, wherein Random Forest exhibits superior performance with MAE of 0.631. The model was trained and tuned in a nested cross-validation.

Furthermore, a comprehensive evaluation of the model's performance is illustrated in Figure 4. The plot demonstrates that a majority of data points exhibit a scattered distribution, not following the red line. Nevertheless, within the grade range of 6 to 8, it can be observed that certain points are more densely clustered in proximity to the line, indicating enhanced predictive capability of the model in this particular range. Conversely, for grades below 6 and above 7.5, the predictions significantly deviate from the expected linear pattern. Additionally, R-square value for the Random Forest model is 0.23, indicating that around 77% of the variability in the dependent variable remains unexplained by the model.

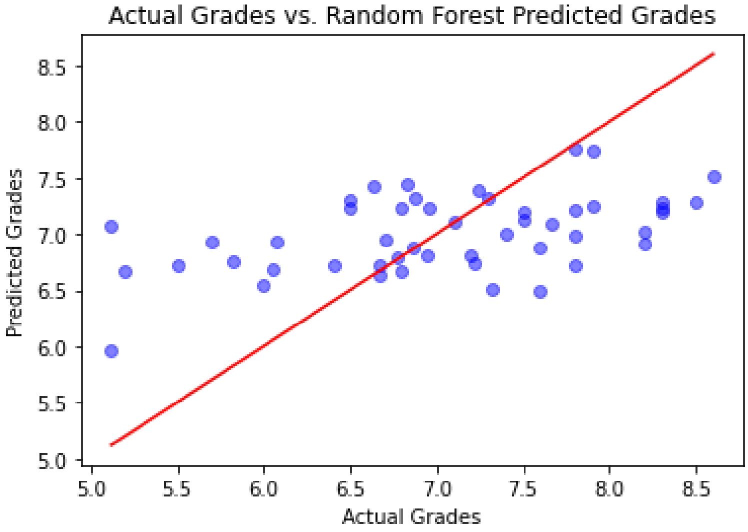


Figure 4: Scatter plot illustration of the predicted and expected values for Random Forest

Figure 5 illustrates a more detailed examination of the distribution of the ground truth and the predicted values. The central tendency for both the true values and predicted values is within the range of 7 to 7.5. The predicted values partly capture the central tendency of the distribution of the target. Nonetheless, it can be seen that the model overestimates grades encompassing 6.5 and 7. This observation aligns with the findings obtained from the scatter plot, further corroborating that the model fails to predict less represented values.

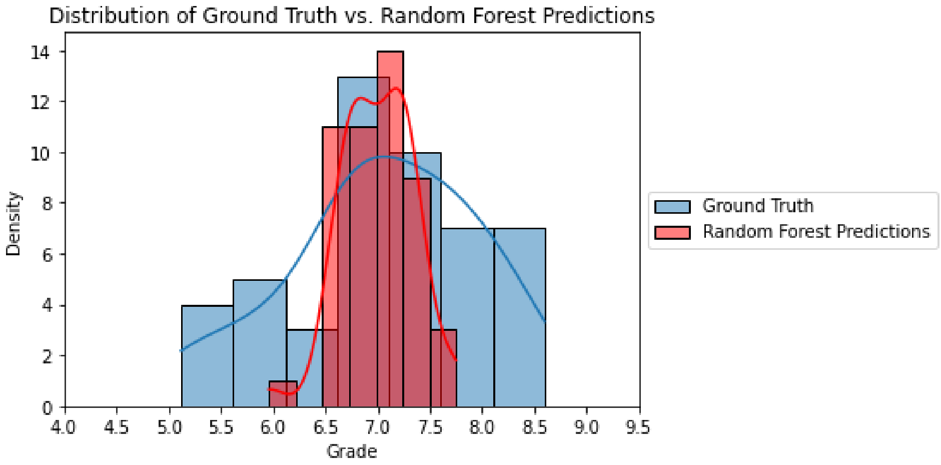


Figure 5: Histogram plot comparing the ground truth GPA and the predicted GPA from Random Forest

### 4.1.2 *Performance of Support Vector Regression*

Support vector regression scored MAE of 0.639 on the test set. This position the model on the second-best predictive performance after Random Forest. The scatter plot in Figure 6 shows that not all the predictions are closely distributed to the red line. The distribution of the dots is similar to the one of Random Forest predictions. Again, the points between grades 6.5 and 8 are more densely and closely distributed to the red line, exposing an acceptable performance in predicting the grades in this group. The R-squared coefficient is 0.22. That shows the model has similar variability to the Random Forest.

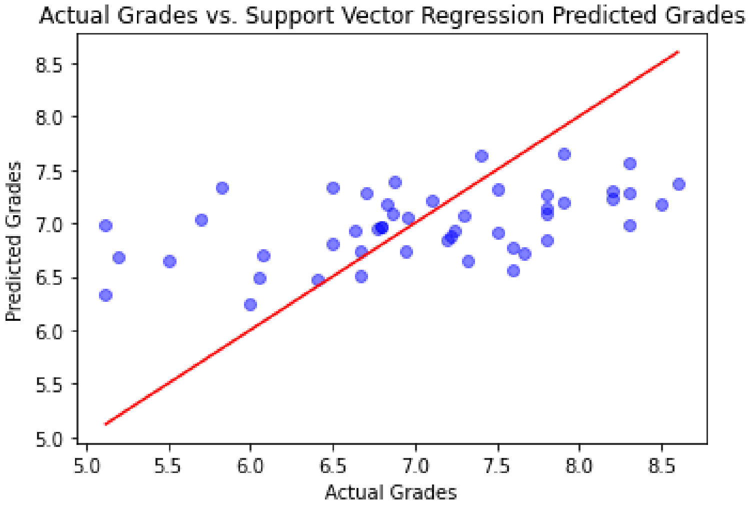


Figure 6: Scatter plot illustration of the predicted and expected values for Support Vector Regression

Figure 7 illustrates the distribution plot, which reveals that the predicted values exhibit a comparable central tendency to the true values. However, a consistent trend emerges where the model tends to overestimate values in the vicinity of 6.5. Additionally, fails to capture and predict values below 6 and above 7.5. This discrepancy suggests that the model performs not-so-good in capturing and predicting instances falling within these particular ranges.

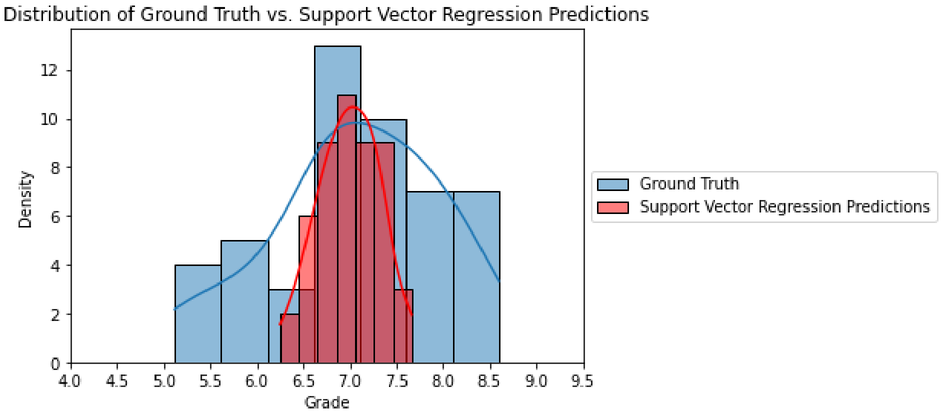


Figure 7: Histogram plot comparing the ground truth GPA and the predicted GPA from Support Vector Regression

### 4.1.3 *Decision Tree performance*

Decision Tree reports 0.65 mean absolute error on the test set. The plot in Figure 8 provides evidence that the model has some capability to capture certain trends in the data. However, it falls short in capturing the complete complexity of the underlying patterns. Furthermore, the presence of two horizontal lines in the predicted values indicates that the Decision Tree model tends to overestimate certain predictions. The R-squared value stands at 0.16. In this case, the low R-squared value suggests that the model fails to account for a significant portion of the variability observed in the true values.

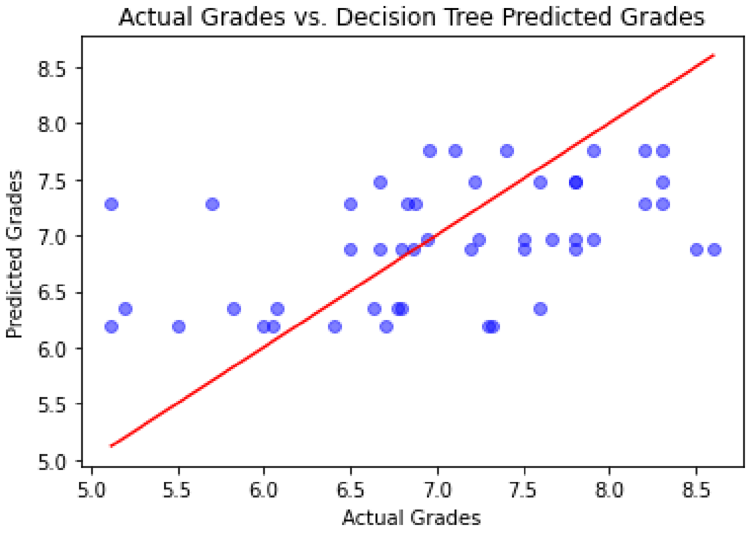


Figure 8: Scatter plot illustration of the predicted and expected values for Decision Tree

From figure 9 it can be seen that the model tends to predict the grades from the majority group. However, overestimation around 6.5 is present in this plot, as well. The separate bars are evident that Decision Tree does not threat some of the predictions as continuous variables.

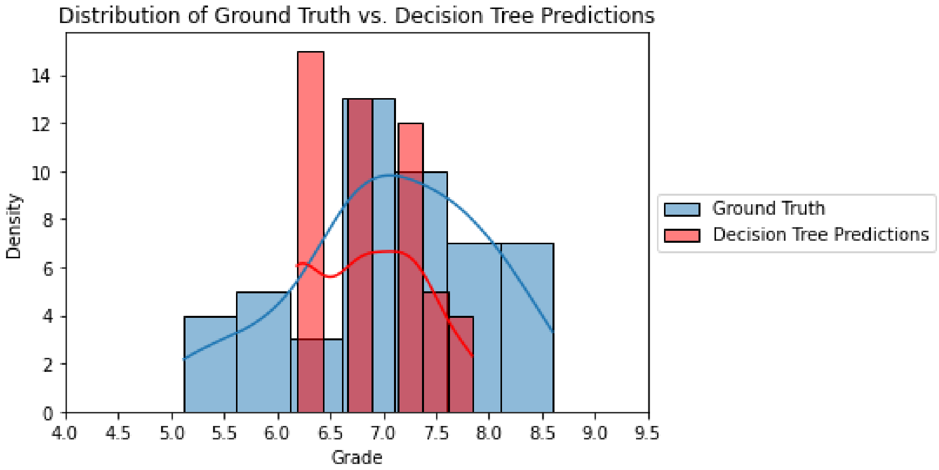


Figure 9: Histogram plot comparing the ground truth GPA and the predicted GPA from Decision Tree

### 4.1.4 *XGBoost performance*

XGBoost exhibited not-so-good performance after applying it on the test set. Furthermore, the model's R-squared score of -0.13 indicates a poor fit to the data. The negative R-squared score suggests that the model fails to capture the underlying patterns and relationships in the data. Figure 10 provides a visual representation of the distribution of points around the diagonal line. Similar to the Decision Tree model, there is a wider spread of data points observed. This indicates a higher degree of variability and inconsistency in the model's predictions, and limitations and shortcomings in the model's predictive capabilities.

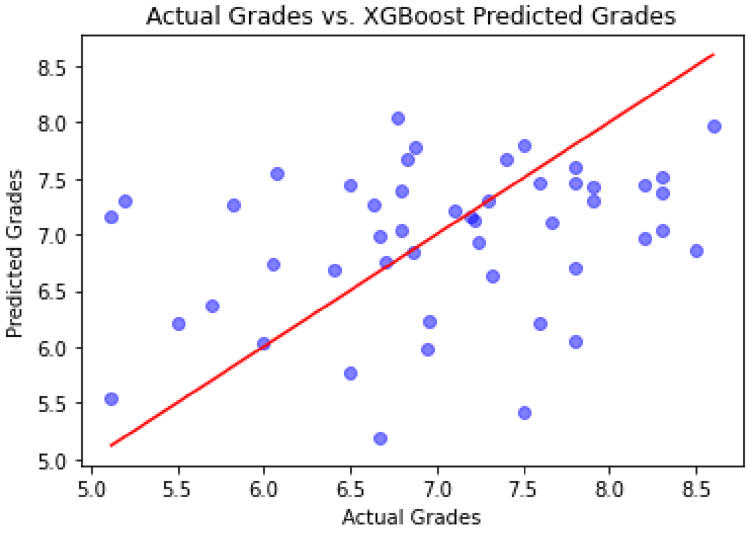


Figure 10: Scatter plot illustration of the predicted and expected values for XGBoost

Interestingly, the histogram (Fig. 11 ) comparing the ground truth with the predicted values, shows that the model is able to capture and predict data from groups with lower observations. This implies that the model's predictions align with the general shape and patterns of the true values. However, it is important to note that capturing the distribution of the data does not necessarily imply accurate predictions or a good fit to the underlying relationships.

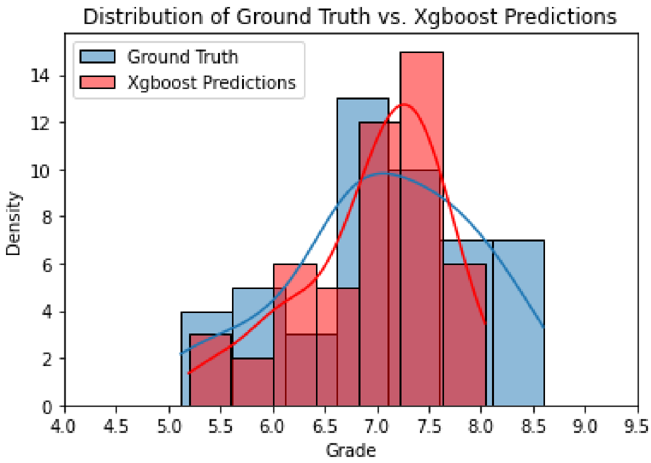


Figure 11: Histogram plot comparing the ground truth GPA and the predicted GPA from XGboost

### 4.1.5 *Linear Regression performance*

Linear Regression does not have any hyperparameters for optimization, therefore the model was only trained with Leave One Out cross-validation. This model ranks last according to it's performance, with MAE of 1.14 and R-squared score of -1.38. Figure 12 shows that the points are widely scattered along the diagonal line, again indicating higher degree of variability and inconsistency in the predictions.

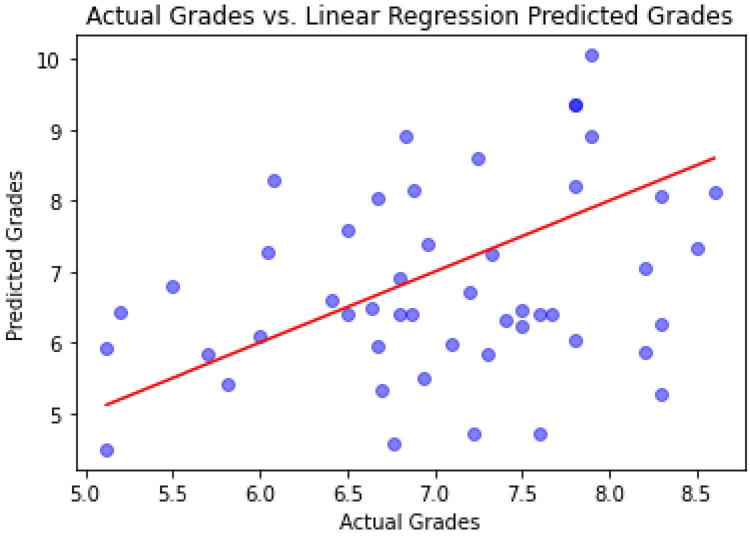


Figure 12: Scatter plot illustration of the predicted and expected values for Linear Regression

The model tends to capture the distribution of the data and some of the central tendency. However, most of the grades are overestimated. Moreover, the model assigns grades that are not present in the test set (Figure 13). These results are not surprising due to the fact that Linear Regression is a simple model that assumes a linear relationship between the predictors and the target. Therefore, it lacks the ability to capture more complex relationships.

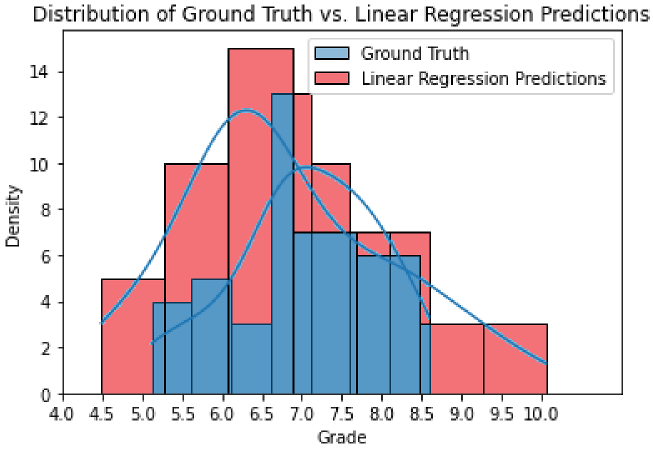


Figure 13: Histogram plot comparing the ground truth GPA and the predicted GPA from Linear Regression

### 4.1.6 *Comparison to the baseline*

The models' performances vary in comparison to the baseline. Random Forest, Decision Tree, and SVR demonstrate superior performance by achieving lower MAE values, indicating their ability to make more accurate predictions. These models exhibit improved predictive capabilities, surpassing the baseline's performance. Notably, Random Forest, Decision Tree, and SVR showcase their effectiveness in handling high-dimensional data, even with 243 observations.

On the other hand, Linear Regression and XGBoost exhibit higher MAE values and negative R-squared values, implying their inferior performance compared to a simple strategy of predicting the mean value. The simplicity of Linear Regression may contribute to its underperformance, as it assumes a linear relationship between predictors and the target variable, which may not adequately capture the complexity of the data. XGBoost, on the other hand, may benefit from further fine-tuning to improve its performance.

## 4.2 *Predictors contribution in the best-performing model*

Figure 14 illustrates the results of the SHAP analysis, which reveals the contribution of features to the predictions made by the Random Forest model. The x-axis represents the average magnitude of the features' influence on the predictions. The top three most influential features in the model's predictions: "Total time spent on campus in January," "Concentration trend slope June," and "Ratio Phone Usage in February in daily window from 12 to 18." These features have demonstrated a substantial impact on the model's predictions. Notably, the majority of the influential features are derived from the app-event dataset and the well-being dataset, encompassing continuous values. This suggests that these engineered features, being continuous in nature, potentially can capture meaningful patterns within the data. Conversely, when considering the one-time dataset containing Likert scale variables and questions related to students' personality, only one feature ("Overcoming difficulties"), was found to have influence in the model's predictions. Overall, the findings suggest that the engineered features derived from the app-event dataset and the well-being dataset, characterized by continuous values, play a role in the Random Forest model's predictions. In contrast, features such as social media usage, instant messaging, streaming services, and internet browsing, which were associated with students' performance in Giunchiglia et al. (2018) study, do not emerge as significant contributors to the final model's predictions.

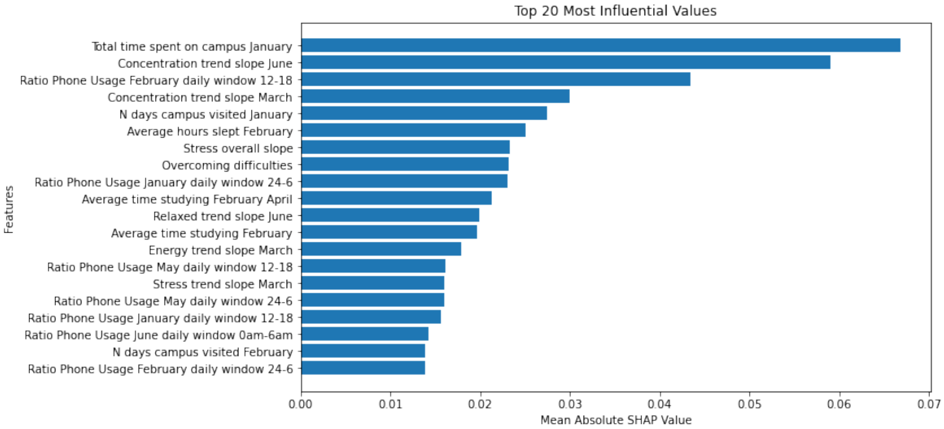


Figure 14: SHAPley values illustrating features contribution to the Random Forest predictions

# 5 DISCUSSION

The research goal of this study was to predict students' GPAs from their mobile usage behaviour and well-being during the semester. The additional sub-research questions were a guideline for finding a solution to this problem. The first research question aimed to determine the machine learning algorithm that can best capture the complexity of the data and provide the most accurate predictions. Subsequently, the objective of the second research question was to find those variables that mostly contribute to the predictions of the best-performing model. The third research question aimed to evaluate the models and their predicting capabilities, using error analysis.

## 5.1 *Results Discussion*

"The analysis of the data suggests that predicting all of the students' grades using the available data is challenging due to limited accuracy. Although some predictions show close proximity to the actual grades, their occurrence is relatively low compared to the total population.

In addressing the first research question regarding the best-performing model for predicting final grades based on the given dataset, Random Forest showed the lowest MAE score of 0.631. This model effectively handled high-dimensional, time-variant data consisting of variables with continuous values that captured various behaviors (such as trend slope, ratio, total time, and total visits).

The findings of this study partially align with existing literature on the subject. Notably, studies conducted by Gadhavi and Patel (2017) and Bai et al. (2021) have also highlighted the predictive capabilities of the Random Forest model. Bai et al. (2021) research showcased the model's robustness when applied to datasets with a larger number of features, encompassing information about patients' behavior inferred from mobile applications. Similarly, Gaftandzhieva et al. (2022) study emphasized the ability of Random Forest to effectively handle time-period datasets spanning a duration of eight weeks.

However, a comprehensive comparison between the studies cannot be conclusively made due to differences in the number of features in the present dataset and variations in the predictors utilized."

In the context of feature contribution to predicting students' GPA using Random Forest, several features were found to have an impact, with contributions ranging from 0.01 to 0.07. While the overall contribution score may appear low, it is important to highlight the significance of specific features in the model. Factors such as university visits, phone usage patterns during daytime (12-18) and evening (18-24), sleep patterns, and certain well-being indicators (e.g., stress, concentration, and relaxation) were observed to contribute to the model's predictions. These findings align with a previous study conducted by Wang et al. (2015), known as the SmartGPA study, which focused on predicting academic performance. In that study, duration-based features and those related to behavioral changes were identified as significant predictors, selected through the Lasso method. Despite methodological differences between the two studies, there is potential for comparison and alignment. Both studies highlight the importance of trend slopes in predicting the final grade, along with the influence of activity during nighttime and evening hours. These similarities suggest that certain aspects of students' behavior and engagement can have an impact on academic performance, providing an opportunity for cross-study insights.

Comparing the models, Random Forest and SVR demonstrate similar performance, with both models achieving relatively lower MAE values compared to Decision Tree and XGBoost. However, Decision Tree and XGBoost exhibit higher errors, suggesting less accurate predictions. These models struggle to capture the underlying patterns and relationships in the data, as reflected in their relatively higher MAE values and negative R-squared score for XGBoost. The scatter plots and distribution plots reveal that all models tend to overestimate grades around 6.5, indicating a common trend. This could be attributed to the limited representation of instances in the training data within this grade range. Overall, while Random Forest and SVR show better performance, there is still room for improvement in accurately predicting students' GPA, particularly for extreme grade values and capturing the full complexity of the underlying relationships.

## 5.2 *Limitations*

One major limitation of this study is the small sample size of the dataset. A larger number of learning observations and a larger test set would increase the likelihood of the models capturing the relationship between the target and the data. This is particularly important when using complex models like XGBoost.

Another limitation is the high dimensionality of the data. Simple models like Linear Regression may struggle to capture the intricate relationships within the data. Incorporating dimensionality reduction techniques could improve the predictions. For example, Wang et al. (2015) and Bai et al. (2021) applied different feature selection methods in their studies, which resulted in higher accuracy. Additionally, exploring different subsets of features and identifying the optimal combination with each model could potentially lead to lower MAE scores.

A possible limitation of the study is the relatively low feature contributions observed in the Shapley values analysis, ranging from 0.04 to 0.07. This suggests that the individual features in the dataset have limited influence on predicting students' grades. Consequently, the overall predictive power of the models may be constrained, and there could be additional unexplored factors or variables that are more significant in determining academic performance, such as inclusion of academic data.

### 5.3 *Contributions and Societal Impact*

This study aimed to contribute to the existing literature by using supervised machine learning techniques to predict academic grades based on students' well-being and behavioral patterns of phone usage. The results indicated that models like Random Forest and SVR demonstrated a higher accuracy in capturing the data that reflects human behavior.

Furthermore, this study highlighted the potential of developing mobile applications that record mobile activity and provide suggestions based on students' past performance. These apps could analyze engagement levels and leisure time spent on the phone to identify if students are at risk of declining grades.

## 6 Conclusion

In conclusion the main research question was answered with the following sub-research questions:

RQ1: Which machine learning model can most accurately capture the relationship between the target variable and the predictors?

The results from Table 3 indicate that Random Forest surpasses the performance of the other models, demonstrating its ability to capture the relationship between students' GPA, their well-being, and inferred phone behaviors. The slight outperformance of Support Vector Regression suggests that this model also holds potential for capturing similar data relationships.

RQ2: What predictors identified by Shapley values contribute to the best-performing model's predictions?

Figure 14 provides an insight on features contribution to the final predictions. The analysis identifies the top three influential features, including "Total time spent on campus in January," "Concentration trend slope in June," and "Ratio Phone Usage in February from 12 to 18 daily window." These features, predominantly derived from the app-event and well-being datasets, exhibit a substantial impact on the model's predictions, emphasizing the significance of continuous engineered features in capturing meaningful patterns within the data.

RQ3: What are the patterns of error across all the models in predicting students' grades?

The scatter plots and distribution plots highlight a common trend of overestimating grades around 6.5, possibly due to limited representation of instances in the training data within this grade range. Although Random Forest and SVR show improved performance, there is still potential for enhancing the accuracy of predicting students' GPA, especially for extreme grade values and capturing the full complexity of the underlying relationships.

## 7 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

The data has been has been provided to me by my supervisor, after signing a statement that I will not disclose any information from the files. The obtained data is anonymised. Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis.

Part of the CODE has been adapted by the author from https://github.com/,https://machinelearningmastery.com/,https://towardsdatascience.com/. The reused/adapted code fragments are clearly indicated in the notebook. In terms of writing, the author used assistance with the language of the paper. A generative language model ChatGPT https://chat.openai.com/ and Gramarly https://www.grammarly.com/ were used to improve the author's original content, for paraphrasing, spell-checking and grammar. No other typesetting tools or services were used. All the graphics and tables were created by the author. Figure 3 was inspired from this illustration https://rb.gy/2tmzo. Data science research pipeline structure was was partly inspired from this figure https://rb.gy/zzm3w.

The codes used in this study can be found on the author's GitHub page https://github.com/MariaNedeva/DSS-Master-Thesis.

## REFERENCES

Abu Saa, A., Al-Emran, M., & Shaalan, K. (2019). Factors affecting students' performance in higher education: A systematic review of predictive data mining techniques. *Technology, Knowledge and Learning*, *24*, 567–598.

Antwarg, L., Miller, R. M., Shapira, B., & Rokach, L. (2019). Explaining anomalies detected by autoencoders using shap. *arXiv preprint arXiv:1903.02407*.

Bai, R., Xiao, L., Guo, Y., Zhu, X., Li, N., Wang, Y., Chen, Q., Feng, L., Wang, Y., Yu, X., et al. (2021). Tracking and monitoring mood stability of patients with major depressive disorder by machine learning models using passive digital data: Prospective naturalistic multicenter study. *JMIR mHealth and uHealth*, *9*(3), e24365.

Batool, S., Rashid, J., Nisar, M. W., Kim, J., Kwon, H.-Y., & Hussain, A. (2023). Educational data mining to predict students' academic performance: A survey study. *Education and Information Technologies*, *28*(1), 905–971.

Chai, T., & Draxler, R. R. (2014). Root mean square error (rmse) or mean absolute error (mae). *Geoscientific model development discussions*, *7*(1), 1525–1534.

El Aissaoui, O., El Alami El Madani, Y., Oughdir, L., Dakkak, A., & El Allioui, Y. (2020). A multiple linear regression-based approach to predict student performance. In *Advanced intelligent systems for sustainable development (ai2sd'2019) volume 1-advanced intelligent systems for education and intelligent learning system* (pp. 9–23). Springer.

Gadhavi, M., & Patel, C. (2017). Student final grade prediction based on linear regression. *Indian J. Comput. Sci. Eng*, *8*(3), 274–279.

Gaftandzhieva, S., Talukder, A., Gohain, N., Hussain, S., Theodorou, P., Salal, Y. K., & Doneva, R. (2022). Exploring online activities to predict the final grade of student. *Mathematics*, *10*(20), 3758.

Gillies, S., et al. (2007–). Shapely: Manipulation and analysis of geometric objects. *toblerity.org*. https://github.com/Toblerity/Shapely

Giunchiglia, F., Zeni, M., Gobbi, E., Bignotti, E., & Bison, I. (2018). Mobile social media usage and academic performance. *Computers in Human Behavior*, *82*, 177–185.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., . . . Oliphant, T. E. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hawi, N. S., & Samaha, M. (2016). To excel or not to excel: Strong evidence on the adverse effect of smartphone addiction on academic performance. *Computers & Education*, *98*, 81–89.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jacobson, N. C., & Chung, Y. J. (2020). Passive sensing of prediction of moment-to-moment depressed mood among undergraduates with clinical levels of depression sample using smartphones. *Sensors*, *20*(12), 3572.

Kotzé, M., & Kleynhans, R. (2013). Psychological well-being and resilience as predictors of first-year students' academic performance. *Journal of psychology in Africa*, *23*(1), 51–59.

Lodder, P., et al. (2013). To impute or not impute: That's the question. *Advising on research methods: Selected topics*, 1–7.

Lundberg, S. M., Erion, G. G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). SHAP: A unified approach to interpreting model predictions [Accessed: June 23, 2023].

Mubarok, F., & Pierewan, A. C. (n.d.). Well-being and academic achievement on students in city of yogyakarta. *Ecopsy*, *7*(1), 374690.

Mukta, M. S. H., Islam, S., Shatabda, S., Ali, M. E., & Zaman, A. (2022). Predicting academic performance: Analysis of students' mental health condition from social media interactions. *Behavioral Sciences*, *12*(4), 87.

pandas development team, T. (2020). *Pandas-dev/pandas: Pandas* (Version latest). Zenodo. https://doi.org/10.5281/zenodo.3509134

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Pratap, A., Atkins, D. C., Renn, B. N., Tanana, M. J., Mooney, S. D., Anguera, J. A., & Areán, P. A. (2019). The accuracy of passive phone sensors in predicting daily mood. *Depression and anxiety*, *36*(1), 72–81.

Rajalaxmi, R., Natesan, P., Krishnamoorthy, N., & Ponni, S. (2019). Regression model for predicting engineering students academic performance. *International Journal of Recent Technology and Engineering*, *7*(6S3), 71–75.

Sinche, S., Hidalgo, P., Fernandes, J. M., Raposo, D., Silva, J. S., Rodrigues, A., Armando, N., & Boavida, F. (2020). Analysis of student academic performance using human-in-the-loop cyber-physical systems. *Telecom*, *1*(1), 18–31.

Van Rossum, G., & Drake Jr, F. L. (1995). *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Wainer, J., & Cawley, G. (2021). Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*, *182*, 115222.

Wang, R., Harari, G., Hao, P., Zhou, X., & Campbell, A. T. (2015). Smartgpa: How smartphones can assess and predict academic performance of college students. *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, 295–306.

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Williamson, R. C., Smola, A. J., & Scholkopf, B. (2001). Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. *IEEE transactions on Information Theory*, *47*(6), 2516–2532.

Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, *98*, 166–173.

Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11.

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: Method review and comparison analysis. *Frontiers in psychology*, *12*, 698490.

# APPENDIX A

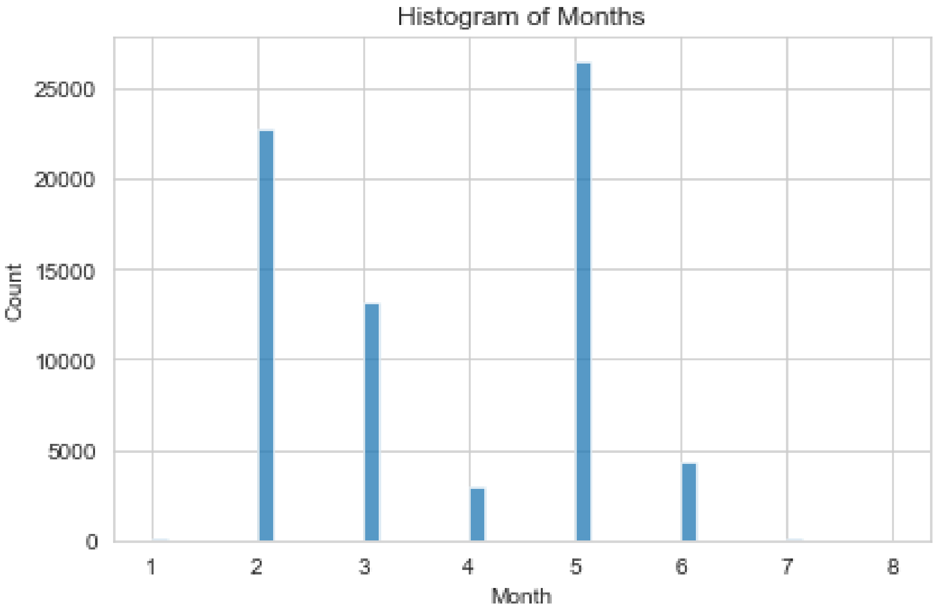


Figure 15: Distribution of the monthly records in the well-being dataset

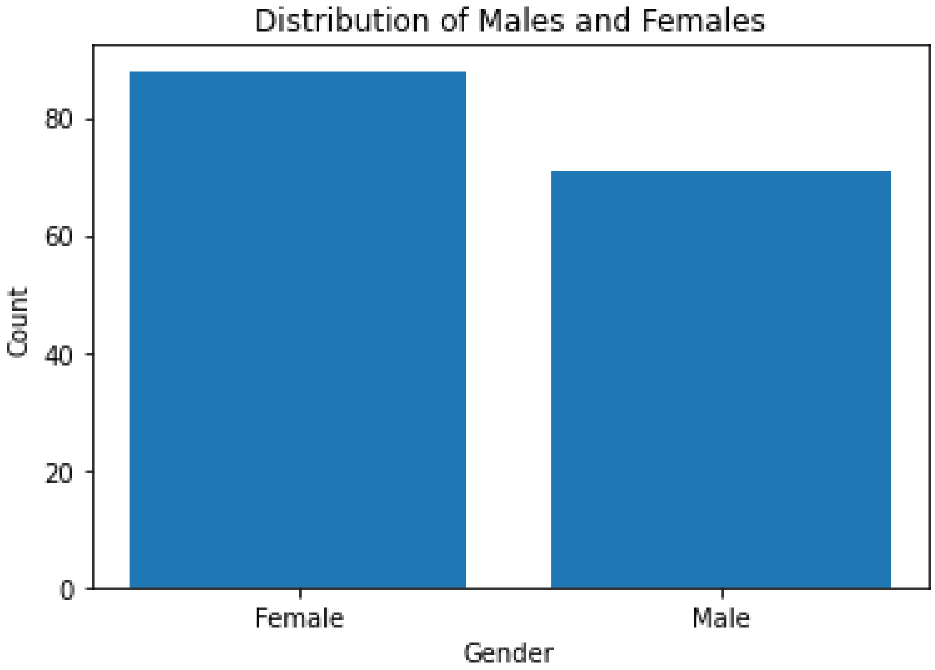


Figure 16: A distribution of gender among students

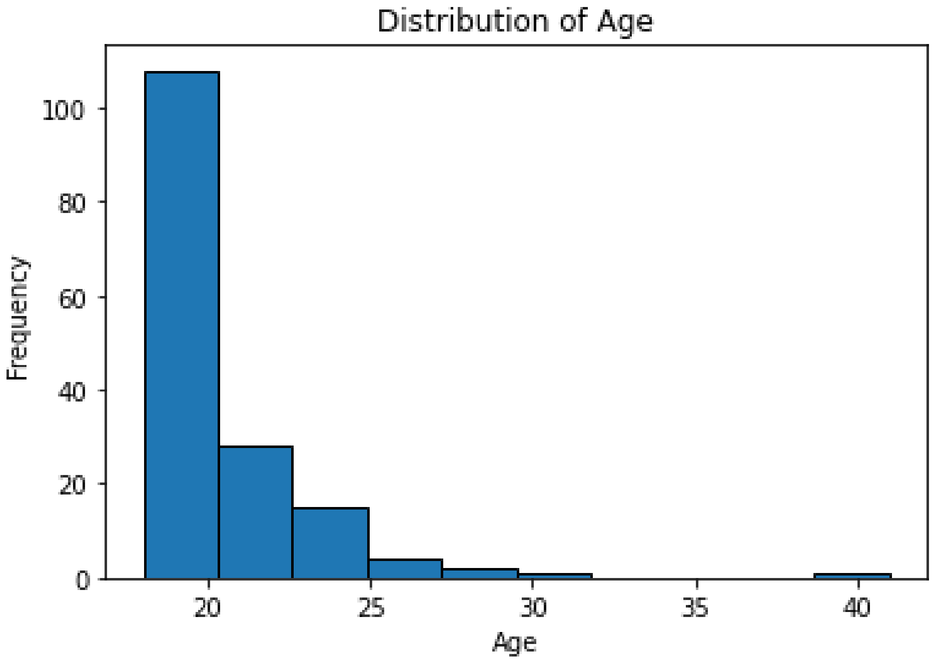


Figure 17: A distribution of age

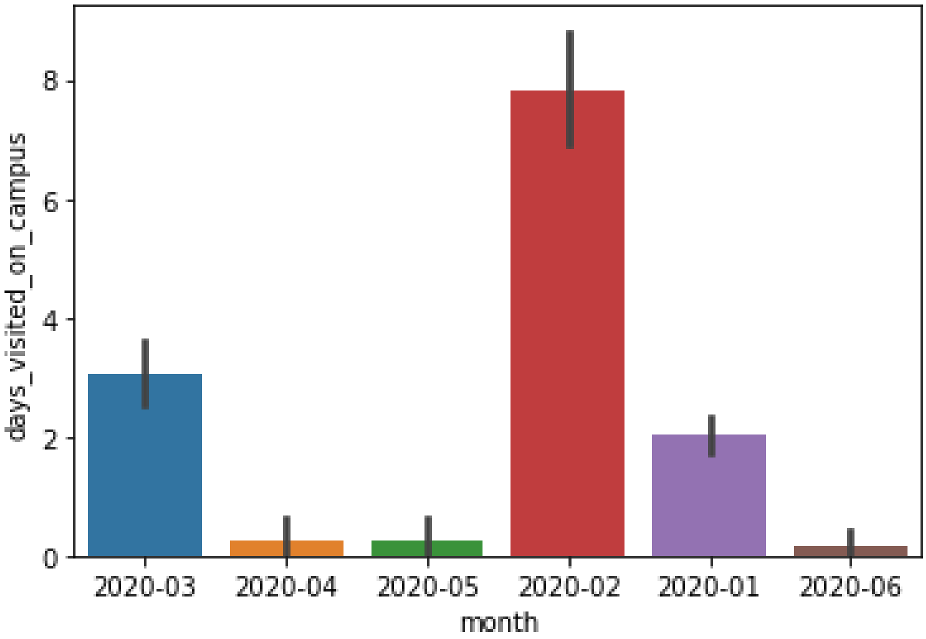


Figure 18: Histogram of frequency of university visits per month

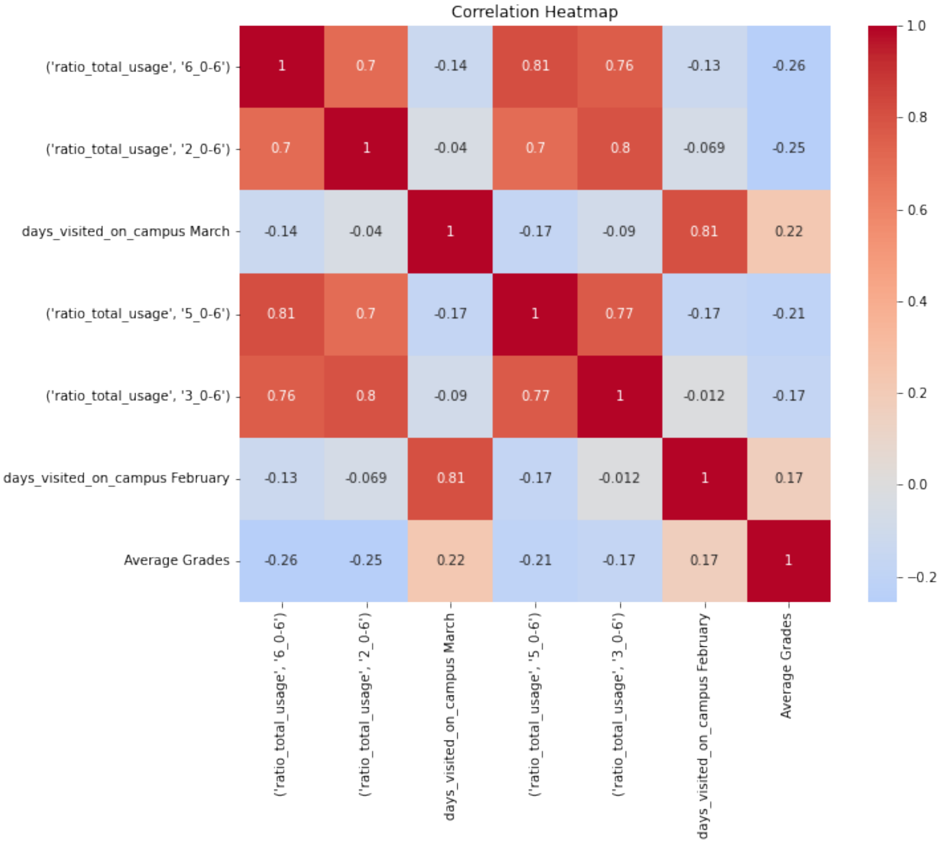
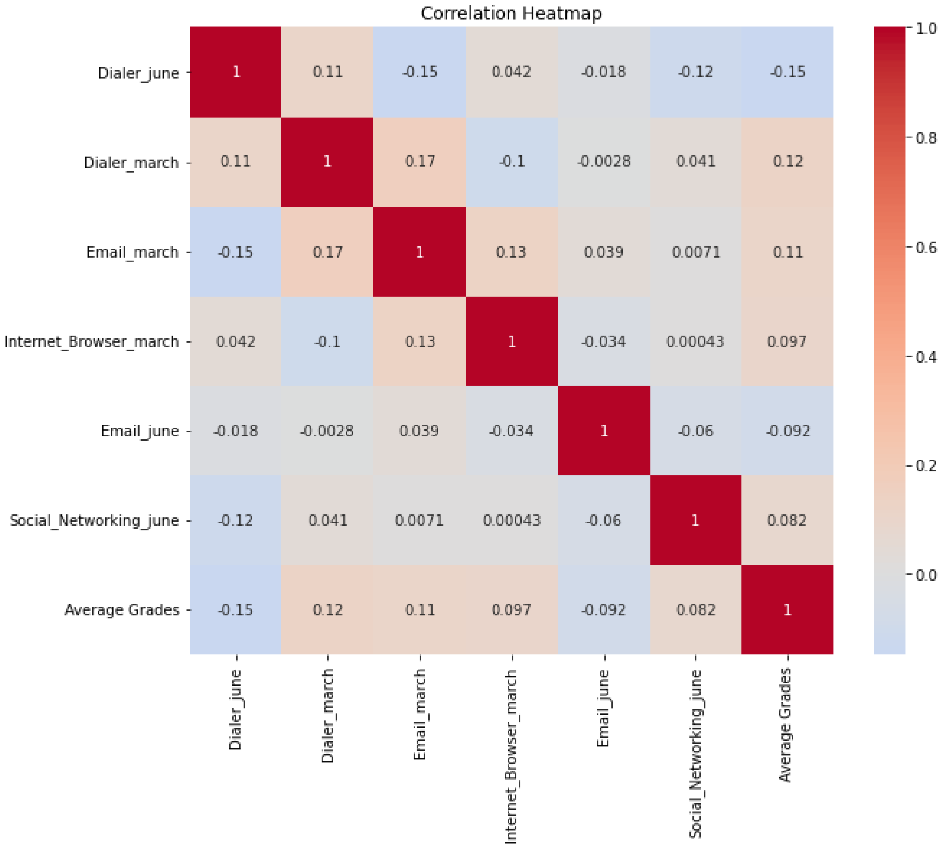


Figure 19: Correlation of Ratio of phone usage, university visits, and average grade. Top 7 most correlated features: 'Ratio of phone usage June from 12 am - 6 am', 'Ratio of phone usage February 12 am - 6 am', 'Total days campus visited March', 'Ratio of phone usage May 12 am - 6 am, Ratio of phone usage March 12 am - 6 am, 'Total days campus visited February', 'Average Grades'



Correlation Heatmap

| | Dialer_june | Dialer_march | Email_march | Internet_Browser_march | Email_june | Social_Networking_june | Average Grades |
|---|---|---|---|---|---|---|---|
| Dialer_june | 1 | 0.11 | -0.15 | 0.042 | -0.018 | -0.12 | -0.15 |
| Dialer_march | 0.11 | 1 | 0.17 | -0.1 | -0.0028 | 0.041 | 0.12 |
| Email_march | -0.15 | 0.17 | 1 | 0.13 | 0.039 | 0.0071 | 0.11 |
| Internet_Browser_march | 0.042 | -0.1 | 0.13 | 1 | -0.034 | 0.00043 | 0.097 |
| Email_june | -0.018 | -0.0028 | 0.039 | -0.034 | 1 | -0.06 | -0.092 |
| Social_Networking_june | -0.12 | 0.041 | 0.0071 | 0.00043 | -0.06 | 1 | 0.082 |
| Average Grades | -0.15 | 0.12 | 0.11 | 0.097 | -0.092 | 0.082 | 1 |

Figure 20: Correlation of applications trend slopes and the average grade

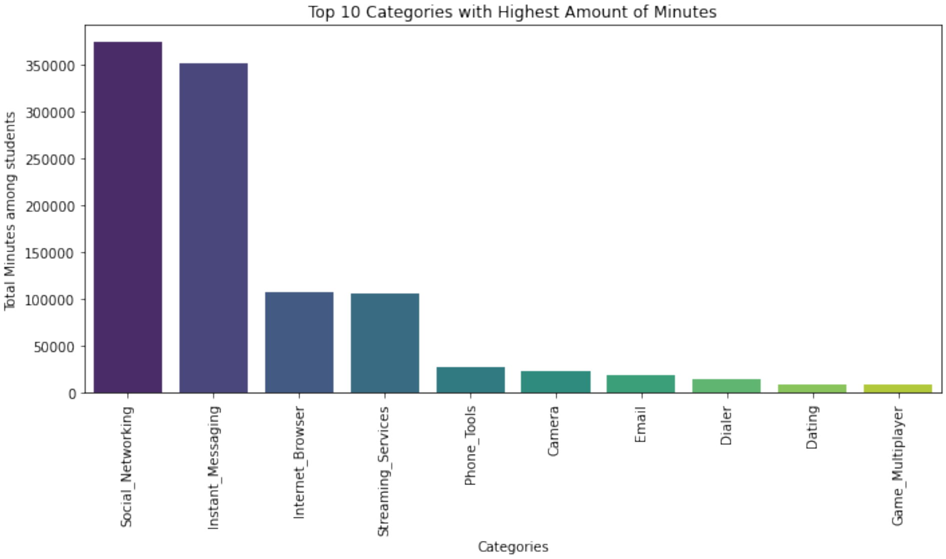


Figure 21: Histogram presenting app categories with spent most minutes on

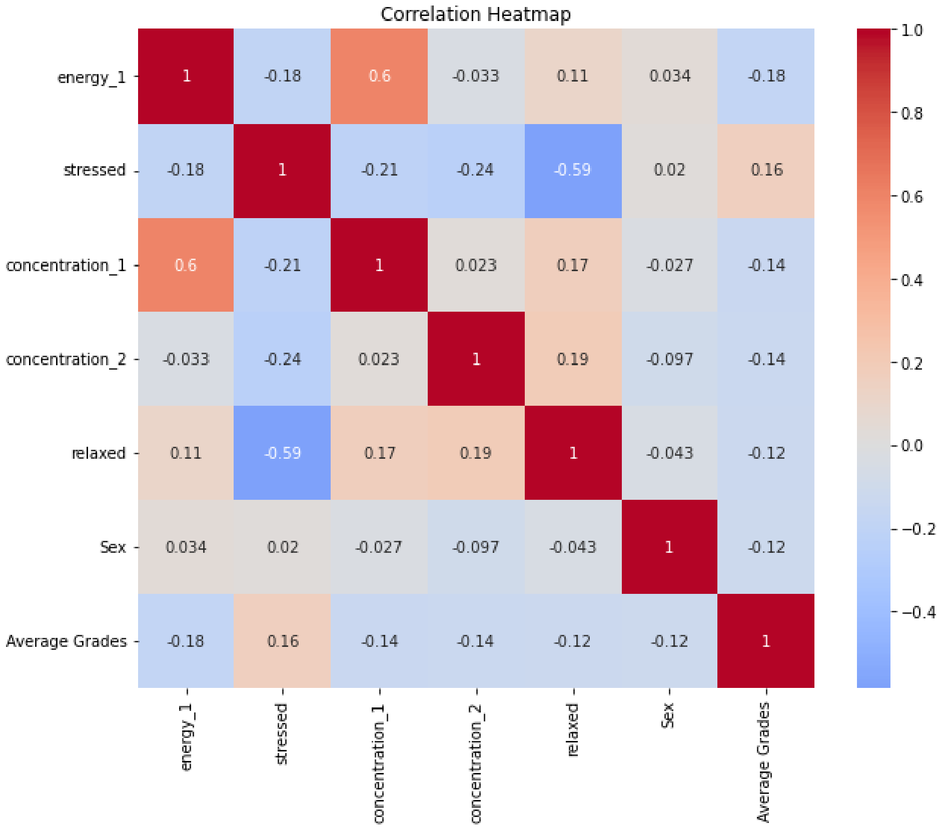


Figure 22: Correlation of well-being slopes with the final grade

## APPENDIX B

Table 4: Support Vector Regression Hyperparameters and Coefficients

| **Hyperparameters** | **Coefficients** |
|---|---|
| Kernel | rbf, poly |
| C | 1, 1.5, 10 |
| Gamma | 1e-7, 1e-4 |
| Epsilon | 0.1,0.2,0.5,0.3 |

Table 5: Decision Tree Hyperparameters and Coefficients

| **Hyperparameters** | **Coefficients** |
|---|---|
| Splitter | best, random |
| Maximum depth | 3, 7, 11, 12 |
| Maximum leaf samples | 1, 2, 3, 4, 5, 6, 7, 8, 9, 10 |
| Minimum weight | 0.1, 0.2 |
| Maximum features | auto, log2, sqrt, None |
| Maximum leaf nodes | 10, 20, 40, 50, 90 |

Table 6: Random Forest Hyperparameters and Coefficients

| **Hyperparameters** | **Coefficients** |
|---|---|
| Splitter | best, random |
| Number of estimators | 100, 200, 300 |
| Maximum depth | None, 5, 10 |
| Minimum Splits | 2, 5, 10 |
| Minimum leafs | 1, 2, 4 |

Table 7: XGBoost Hyperparameters and Coefficients

| **Hyperparameters** | **Coefficients** |
|---|---|
| Splitter | best, random |
| Learning rate | 0.1, 0.3 |
| Maximum depth | 3, 5 |
| Subsample | 0.6, 0.8 |
| Colsample Bytree | 0.6, 0.8 |

Table 8: Optimal Hyperparameters for Each Model

| **Model** | **Optimal Hyperparameters** |
|---|---|
| SVR | C=1, epsilon=0.5, gamma=1e-07, kernel='poly' |
| Decision Tree | Max depth: 7, Max features: 'sqrt', Max leaf nodes: 20, Max samples: 2, Min weight leaf: 0.1 |
| Random Forest | Max depth: 10, Max samples split: 10 |
| XGBoost | Learning rate: 0.3, Colsample: 0.6, Max depth: 3, Estimators: 100 |