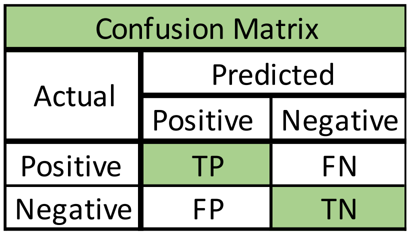
Introduction

Exam grades have been a heavily weighted portion of most college students academic career. Many factors can affect exam scores such as time of day, teaching style, amount of students in the class, and mode of instruction. These factors can be an unfair advantage to students who might prefer small classes, night over morning classes, or a pure online mode of instruction. The biggest factor that can be researched is the time of day the class starts, which is heavily influenced by the amount of sleep students get. Researchers from Saint Lawrence University [1] have shown that the later the class starts the lower the average grade is. Saint Lawrence University observed that for every hour later the class starts from an average national start time of 8:00 A.M. that the average grades will decrease by 0.02 per hour [1]. The researchers observed that the earlier the class time the earlier the student is inclined to sleep earlier compared to the students taking night classes who will stay up later in the night. The University of Michigan [2] found that 75% of undergraduates do not sleep enough to feel rested on five or more days per week with 19% reporting that less sleep had an impact on their academic performance. Also a study done at a Texas University [3] conducted with 829 undergraduate students showed that early bird students had an average GPA of 3.5 compared to night owls with an average GPA of 2.5. These observations and results bring up an interesting question that data science can solve. The question is that if studies have shown that the later the class starts the lower the grades are that if there is a big enough difference it might be possible to classify and predict the class time based solely on the grades. In an attempt to solve this question I have collected exam grades from the undergraduate statistics class STA 2023 Statistical Methods I at the University of Central Florida with the first class starting at 10:30 am until 11:45 am taught by Daniel Inghram, associate instructor, and the second class starting at 4:30 pm until 5:45 pm taught by Lixia Wang, visiting instructor. I was able to obtain this data, since I am a graduate teaching assistant for the course. This class has a diverse student background from freshman to senior. So the purpose of this study is to conduct a classification and prediction analysis using Naïve Bayes, Logistic Regression, and Support Vector Machines on the two classes afternoon and morning and then calculating various performance metrics such as precision, accuracy, sensitivity, specificity, F1 Score, Negative Predicted Value, False Negative Rate, and False Positive Rate.

Methods

Before I begin the classification analysis I had to collect the data from the morning and afternoon sections. Since I am a graduate teaching assistant for the morning section at interest I had instant access to the students Exam 1 and 2 grades. For the afternoon section my colleague who is the graduate teaching assistant for the afternoon section of interest was willing to extract the Exam 1 and 2 grades with all identifiable information removed such as student name, UCF ID, and email due to FERPA regulations. There are other grade information such as iClicker questions and online homework, but these are not reflective of the students actual class performance, since they are done with access to the internet so they are omitted from this analysis. Some exam 1 and 2 grades had to be omitted from the morning section, since the class has more students so the sample size was set to 329 exam grades per section. To validate the past research done by Saint Lawrence University [1] I did a 5-point descriptive statistics calculating the mean, standard deviation, median, minimum value, and maximum value of each section to ensure that there is a difference in exam grades. I then assigned labels being afternoon or morning, which is the response variable for the classification. I set two predictors for the classification being Exam 1 and Exam 2 grades so the total data set dimension was 658 x 3 (first 329 afternoon, second 329 morning). To perform the classification and prediction a training and testing set have to be randomly created so I decided for 80% of the data for training and 20% for testing. For classification and prediction I decided to use Naïve Bayes, Logistic Regression, and Support Vector Machines with a linear kernel, because these methods are optimal for binary outputs. Naïve Bayes is a conditional probability approach based on Bayes Rule where a Bayes Decision Boundary can be drawn to identify different classes. Logistic regression uses regression to perform the classification based on the logit function. Support Vector Machines uses a hyperplane approach to the separate out the data based on class using various kernel functions such as linear and radial basis function kernels. The classification and prediction was done using R with the following packages: readr, naïvebayes, nnet, ISLR, e1071. Once the classification and prediction was performed I generated confusion matrices, using the R package caret, per each classification method using the 20% testing data (Actual) and the fitted predicted values (Predicted).



Confusion Matrix		
Actual	Predicted	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Figure 1 General 2x2 Confusion Matrix

For the confusion matrix the positive class is set as the afternoon class, which means it is bad, since the afternoon classes receive lower grades as shown by past research [1,2,3]. A True Positive (TP) represents a correctly classified positive observation. A False Positive (FP) represents a falsely classified positive observation. A False Negative (FN) represents a falsely classified negative observation. A True Negative (TN) represents a correctly classified negative observation. From the confusion matrix I calculated eight performance metrics being:

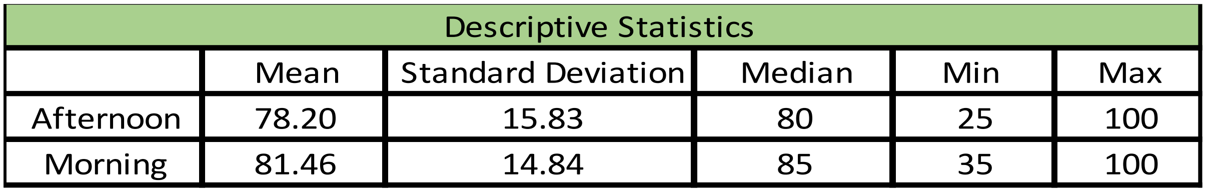
$$\text{Accuracy} = \frac{TP+FN}{TP+FP+FN+TN} \qquad \text{F1 Score} = 2 * \frac{\text{Precision x Sensitivity}}{\text{Precision +Sensitivity}} \qquad \text{Specificity} = \frac{TN}{TN+FP}$$

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Negative Predictive Value (NPV)} = \frac{TN}{TN+FN} \qquad \text{False Positive Rate (FPR)} = \frac{FP}{FP+TN}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \qquad \text{False Negative Rate (FNR)} = \frac{FN}{FN+TP}$$

The accuracy measures how accurate the model is classifying the actual classes. The precision or positive predicted value (PPV) measures the positive classification performance. The sensitivity measures how well the positive class are correctly identified. The F1 score represents the weighted average of precision and sensitivity. The specificity measures how well the positive negative class are correctly identified. The NPV measures the true negative classification performance. The FNR measures the false negative classification performance. The FPR measures the false positive classification performance. For accuracy, precision, sensitivity, specificity, F1 score, and NPV the closer the value is to 1 the better the model is. For FNR and FPR the closer the value is to 0 the better the model is. From the performance metrics future recommendations and conclusions are discussed.

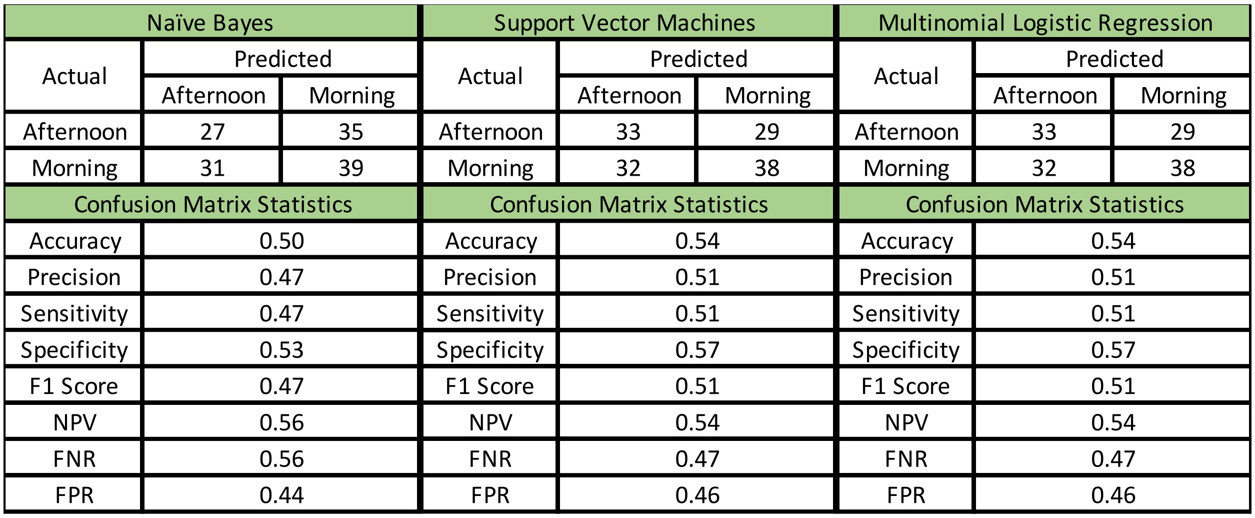
Results



Descriptive Statistics					
	Mean	Standard Deviation	Median	Min	Max
Afternoon	78.20	15.83	80	25	100
Morning	81.46	14.84	85	35	100

Figure 2 5-Point Descriptive Statistics

As shown in in figure 2 we can observe that the mean exam score for the morning section is greater by 3.26% and has a higher minimum grade. These findings validate the past research [1,3] that having an earlier class start time is correlated to higher overall grades.



Naïve Bayes			Support Vector Machines			Multinomial Logistic Regression		
Actual	Predicted		Actual	Predicted		Actual	Predicted	
	Afternoon	Morning		Afternoon	Morning		Afternoon	Morning
Afternoon	27	35	Afternoon	33	29	Afternoon	33	29
Morning	31	39	Morning	32	38	Morning	32	38
Confusion Matrix Statistics			Confusion Matrix Statistics			Confusion Matrix Statistics		
Accuracy	0.50		Accuracy	0.54		Accuracy	0.54	
Precision	0.47		Precision	0.51		Precision	0.51	
Sensitivity	0.47		Sensitivity	0.51		Sensitivity	0.51	
Specificity	0.53		Specificity	0.57		Specificity	0.57	
F1 Score	0.47		F1 Score	0.51		F1 Score	0.51	
NPV	0.56		NPV	0.54		NPV	0.54	
FNR	0.56		FNR	0.47		FNR	0.47	
FPR	0.44		FPR	0.46		FPR	0.46	

Figure 3 Summary of Performance Metrics

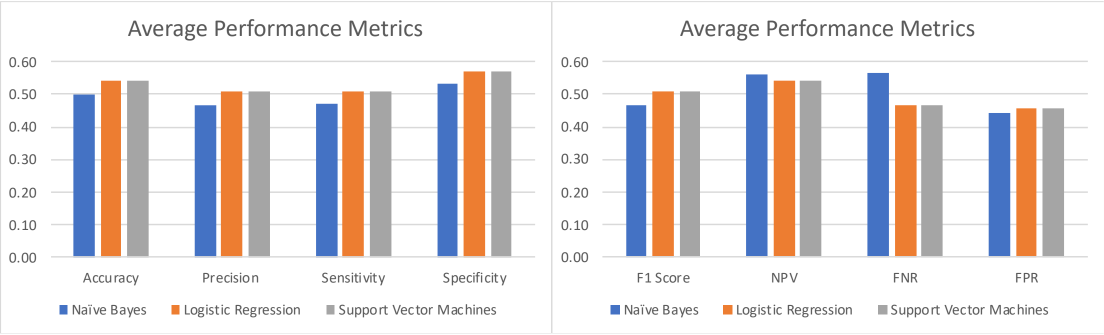


Figure 4 Visualization of Performance Metrics

Looking at figure 3 for the summary of performance metrics we see that the average performance for accuracy, precision, sensitivity, specificity, F1 Score, PPV, and NPV are around 0.50, which is not as well as I expected since 1.0 represents the best performance. For the FNR and FPR the average was also around 0.50, which is not as well as I expected since the 0.0 represents the best performance.

Discussion

We can see that from the classification and prediction analysis that it is possible to distinguish between morning and afternoon exam grades with a sample size of 329 undergraduate students, but with very low performance. But with the 5 point descriptive statistics I was able to show that the mean exam grade for the morning section was greater by 3.26%, which validates the studies done at Saint Lawrence University [1] and the university from Texas [3] that the earlier the start time for a class the higher the overall grades are. A factor I did not consider for this classification analysis is the amount of teaching experience that Daniel Inghram, associate instructor, and Lixia Wang, visiting instructor, for the Statistical Methods I course, since Daniel Inghram has been with the University of Central Florida for four years now compared to Lixia Wang who is on a visiting instructor assignment. For the professor to be a considered a factor in average exam grades I would have to collect more data with Daniel Inghram teaching the class later and Lixia Wang teaching the class early, and then I can see if the course professor has an impact on the average exam grades.

Future Work & Conclusions

To better improve the classification performance metrics I can collect more data from other sections of STA 2023 Statistical Methods I, but the earliest the class is offered is 10:30 am and the latest being 4:30 pm so it would be up to the University of Central Florida to change the future class times to see if the overall exam grade average across the sections increases with earlier start times such as 8 am. Also, I can add more predictor variables such as the online homework, in class iClicker grades, and the average amount of sleep that each section gets per school night to see if feeding more data into the classifiers with improve the classification performance. Also, collecting solely exam grade data from all sections over a set amount of years might help improve the classification since every fall semester around 1500 undergraduate students take STA 2023 Statistical Methods I and the total sample size for my classification was 658 undergraduate students (329 undergraduate students per section). So having a larger sample size might help detect a trend, since I was able to obtain an average of 50% classification performance so increasing the sample size to $n > 10,000$ might increase the classification performance. To also improve the classification on the smaller sample size I can try other classification methods that can be tuned such as Support Vector Machines with a radial basis function kernel with parameters cost and gamma, Bagging, Boosting, Random Forest were you can vary the amount of classification trees, and k nearest neighbors to compare them to the Naïve Bayes, Logistic Regression, and Support Vector Machines performance metrics. The last way to improve classification performance would be to vary the training and testing split, but it might end up decreasing the classification performance. So overall it is possible to classify and predict exam grades with a sample size of 629 undergraduate students with 80% set for training and 20% for testing, but with low performance.

References

1. Reimold, Dan. “Early Classes Equal Higher College Grades, Study Confirms.” Usatoday.com, Usatoday, 30 Oct. 2011.

2. “Successful Students Tend to Sleep More.” Umich.edu, University of Michigan, 2019.

3. Laino, Charlene. “Early Birds Get Better Grades.” Webmd.com, Webmd, 9 June 2008.

Appendix A: Supplementary R Code

```
#import combined dataset with labels
library(readr)
combined <- read_csv("~/Google Drive/Graduate College Docs/PhD/Fall 2019/SAS/Project/combined.csv")
smp_size <- floor(0.80 * nrow(combined))

## set the seed to make partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(combined)), size = smp_size)
train <- combined[train_ind, ]
testing <- combined[-train_ind, ]
train = as.data.frame(train);
testing = as.data.frame(testing);

##Naive Bayes
library(naivebayes)
train$Section = as.factor(train$Section)
model.nb.train <- naive_bayes(Section ~.,data=train[,1:3])
nb.pred.testing <- predict(model.nb.train, testing[,1:2], type="class")#Get the fitted labels

###Logistic Regression
library(nnet)
train$Section = as.factor(train$Section)
model.mlr.train <- glm(Section ~.,family=binomial(link='logit'),data=train[,1:3])
train.multinom <- multinom(Section ~., data=train, maxit = 1000) # Fitting
testing.multinom.pred <- predict(train.multinom, testing[,1:2] ,type="class")

##Support Vector Machines
library(ISLR)
library(e1071)
train$Section = as.factor(train$Section)
svm_linear.model <- svm(Section ~ ., data=train , kernel ="linear",cost=.01)
svm_linear.pred <- predict(svm_linear.model, testing[,1:2])

#getting the confusion matrices
library(caret)
naivebayes_matrix <- confusionMatrix(table(testing[,3], nb.pred.testing))
svm_matrix <- confusionMatrix(table(testing[,3], svm_linear.pred))
mlr_matrix <- confusionMatrix(table(testing[,3], testing.multinom.pred))
```