Partners: Lana Huang lh711, Sharon Chen sc2351

CS439 Final Project Report

Video | GitHub

Project Definition:

1.1 Problem Statement

The rise of social media and the growing number of youth engaged digitally through social media have resulted in a very heavy impact on adolescent daily life and mental well-being. The average young adult in the U.S. alone spends at least 3 hours a day on social media, which is associated with increasing levels of anxiety, depression, and addiction (eMarketer 1). Such increases in screen time and accompanying decline in mental well-being represent one of the many ongoing developments within our currently online-driven society. Beyond individual well-being, these behavioral patterns may expose young adults to greater cybersecurity risks–ranging from privacy breaches to phishing attempts–raising the question of whether national trends in digital behavior correlate with broader cybersecurity vulnerabilities. In our project, we investigated how certain types of behaviors by younger generations in relation to social media, mental health, and addiction may affect cybersecurity scores, specifically the cybersecurity exposure index (CEI), within countries.

1.2 Connection to Course Material

This project is connected to several key topics in our course: CS439 Introduction to Data Science. This includes data collection and management by importing, cleaning, and combining datasets. Through the use of pandas and the matplotlib library, which we go over in class, we can manipulate data, do exploratory data analysis (EDA), and create visualizations of our data relationships. Our project will also demonstrate the application of statistical techniques, such as using correlation analysis to determine correlation coefficients and significance, which we will explore alongside machine learning techniques, such as linear regression and random forest regression models.

Novelty and Importance:

2.1 Importance of the Project

Our project is important because now that we're in the age of technology, the technology is going to be constantly evolving and getting better, but with that, security risks are getting bigger and bigger. They also get more complex and harder to safeguard around them. Not only does it affect many corporations and organizations, but it also affects normal individuals. For instance, individuals have to protect their personal data, like address, identity, and financial information. As a result of a breach, they can have their bank information stolen or even their identity and social security number. For corporations, there are risks that include financial data for their clients or their own business. They also have issues where cyber attacks can result in a shutdown of a service, as we've seen most recently in the AWS outage, where many online services were not accessible.

2.2 Excitement and Relevance

We are excited to learn more about this relationship because both of our lives are intertwined with our phones and online spaces. It's something that we enjoy, and to be able to keep us safe is important as we navigate an online world. We are also excited to learn more about cybersecurity, as it is a future career path that both of us are interested in. We want to see how cybersecurity can be integrated into our daily lives, like using our phones and being safe online, and develop good practices for navigating online spaces.

Progress and Contribution:

3.1 Data Utilization:

For our first dataset, we wanted to look at the relationships within each dataset before focusing on the cross-dataset relationships. We first looked at the student mental health dataset, where we found relationships that we were expecting. In Fig.1, we can see that the relationship between social media usage and the mental health score trends downwards. This is what we expected, as the higher the social media usage, the lower the mental health score. Then, we can see in our Fig. 2, we mapped sleep hours per night against the mental health score. This also matched our expectation, as the more hours of sleep at night, the higher the mental health score. Then, in Fig. 3, we see that as we map addiction across mental health that the more addicted students feel towards social media, the lower the mental health score.

Our second dataset revolves around the Global Cybersecurity Index for 192 countries. We used this data to track primarily only these features: “Country”, “Region”, “CEI”, and “GCI”. CEI refers to the Cybersecurity Exposure Index, the higher the index, the higher the vulnerability to cybersecurity threats. GCI refers to the Global Cybersecurity Index, which measures a country's cybersecurity advancements. Countries with high GCI usually have strong cybersecurity policies, defense, and infrastructure (Wang, Hu, et al. 83). Through these, we mapped the relationship between CEI and GCI, which was highly interesting because it showed us an almost parabolic downward trend. The higher the GCI scores, the lower the CEI values, which makes sense. Countries with strong cybersecurity precautions have a lower vulnerability to cybersecurity threats.

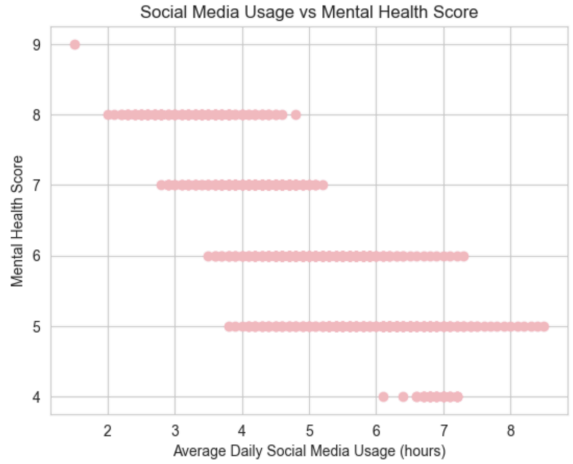


Fig.1

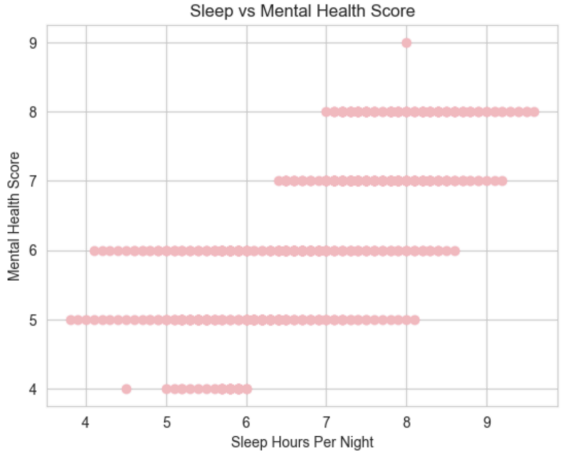
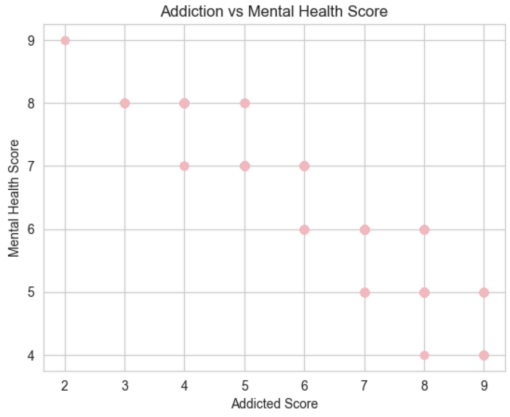


Fig. 2



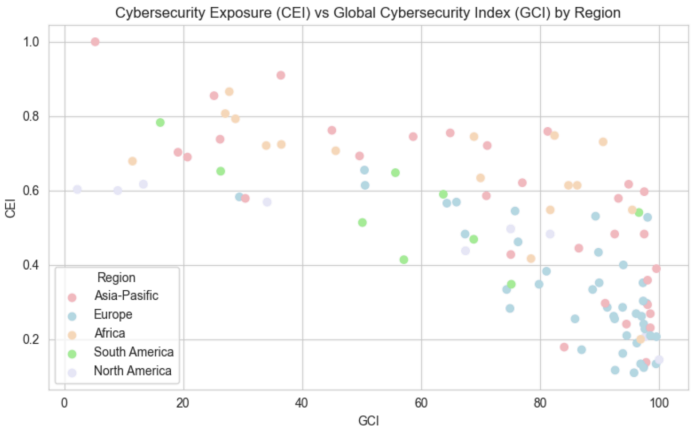


Fig. 3 Fig. 4

These datasets are exactly what we need in order to map the relationship between the measure of cybersecurity vulnerability and that country's student body mental health and well-being.

3.2 Models, Techniques, and Algorithms

In this section, we will talk about the Models used for our project, but we will discuss the evaluations and analysis for later in the report.

For our project, we wanted to examine how students' mental health and online activity could relate to their country's CEI index. For mental health and online activity, we used the feature: "Mental Health Score". We started with the traditional Linear Regression, which was our first test with the Regression model, in order to see the relationship between young adult Mental Health Score and a country's CEI. We wanted to start with Linear Regression because it's the simplest possible regression model that we could run, and we wanted to start basic before enhancing the model. We split the data into the normal 80% training set and the 20% testing set in order to prevent overfitting. We then use StandardScaler() to standardize the predictor. We then train the model and make predictions.

Next, we wanted to try a different model type rather than Linear. Instead, we decided to test Random Forest Regression because it handles non-linear patterns that linear models cannot. Since our linear regression model didn't perform well, we realized the relationship might not be linear but, in fact, more complex. The single variable random forest regression model didn't end up performing well either, but we decided to stay with Random Forest. As seen in fig. 5, we have the number of decision trees at 250. Originally, we had it at 100 to start with, but as we increased it, we realized that it had improved the model even if only by a little bit. This makes sense as more trees are able to better at generalizing.

```
# Create and train the regression model
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

model = RandomForestRegressor(n_estimators=250, random_state=42)
model.fit(X_train_scaled, y_train)
```

Fig. 5

Instead, we decided that we should include not just one predictor variable, but actually, we can use multiple. Rather than just using the "Mental Health Score", we also used "Average Daily Screen Usage", "Sleep Hours per Night", and "Addiction Score". Now our question has shifted from predicting CEI using mental health scores to predicting CEI using a combination of mental health indicators. Now, in this one, the output was the best compared to the other two models. For this one, we also made sure to play with the n_estimators to get the best possible model without sacrificing training time for a little bit of improvement.

```
#Create and train Multi-Variable Random Tree Model
X_multi = merged_df[['Mental_Health_Score', 'Avg_Daily_Usage_Hours', 'Sleep_Hours_Per_Night', 'Addicted_Score']]
y_multi = merged_df['CEI']

X_train_multi, X_test_multi, y_train_multi, y_test_multi = train_test_split(
    X_multi, y_multi, test_size=0.2, random_state=42
)
✓ [39] < 10 ms
```

```
#Scale and train random Forest Regressor
scaler_multi = StandardScaler()
X_train_multi_scaled = scaler_multi.fit_transform(X_train_multi)
X_test_multi_scaled = scaler_multi.transform(X_test_multi)

model_multi = RandomForestRegressor(n_estimators=300, random_state=42)
model_multi.fit(X_train_multi_scaled, y_train_multi)
✓ [40] 156ms
```

Fig. 6

3.3 Experimental Design

Our goal for this project is to systematically test how mental health and digital behavior variables could predict cybersecurity exposure (CEI). To do this, we first started with a student social media dataset, which included the country a student was from and their own digital behaviors and mental health status (“Mental Health Score”, “Average Daily Screen Usage”, “Sleep Hours per Night”, and “Addiction Score”). We then had a cybersecurity dataset that compiled countries’ cybersecurity KPI’s like CEI and GCI. We made sure to merge the dataset on the country column to be our key. Then we preprocessed our data using standard cleaning practices by dropping columns, imputing missing data, and finding and correcting outliers. Then we selected the different models. We started with Linear Regression to start our model off simple as a baseline before moving on to more complex models. Then we moved to a single variable random forest model where we only used “Mental Health Score” as a predictor. As we fine-tuned our model, we decided to move on to adding more predictors to the random forest model in order to improve prediction accuracy. For each model, we made sure to split and train the data according to how we were taught. We then evaluated the models by looking at their R-squared, mean squared error, and the mean absolute error. We decided to go by these errors because R-squared tells us how much variance is explained, MSE penalizes larger errors, and MAE shows the average prediction error. We then created an interactive slider similar to the one from lab 3 in order to show how each predictor changes the CEI index.

3.4 Contribution

As a team, we decided from the start to practice pair programming. Since we live together, we were able to meet at least once a day to discuss the project, goals, and code. We also decided that this would be the best way to work on it because of our living situation, and that it would be easier to collaborate instead of delegating tasks. We didn’t want to be cut off from participating in different tasks and practices. Due to this setup, we made sure to collaborate on each code implementation, trading off and on for who's the driver and who's the navigator. We switched roles after each section and started with Sharon as the first driver and Lana as the navigator. We believed this was the best scenario to enforce equal work for this project.

Detailed Analysis:

4.1 Key Findings and Results

For our project, we trained 3 different regression models to study the relationship between mental health and the national cybersecurity exposure index: a linear regression model, a single variable random forest regression model, and a multi-variable random forest regression model.

```
Linear Regression Model
R-squared Score: 0.2151
Mean Squared Error: 0.0343
Mean Absolute Error: 0.1458
```

Starting with our linear regression model, we were able to predict the CEI regression values from mental health scores. We found that our regression model yielded a R^2 score of 0.2151, explaining 21.51% of the variance in CEI, with a mean squared error (MSE) of 0.0343, and a mean absolute error (MAE) of 0.1458. While this shows a measurable linear relationship, we wanted to explore more models to see if we could obtain improved values.

```
Single Variable Random Forest Regression Model
R-squared Score: 0.2104
Mean Squared Error: 0.0345
Mean Absolute Error: 0.1469
```

Seeking improvement from our linear regression model, we moved onto using a single variable random forest regression model to predict the CEI regression values from mental health scores. We found that this regression model yielded a R^2 score of 0.2104, explaining 21.04% of the variance in CEI, with a mean squared error (MSE) of 0.0345, and a mean absolute error (MAE) of 0.1469. Unexpectedly our single variable regression model showed lower R^2 score and higher MSE and MAE scores than the linear regression model, with a 0.0047 decrease in R^2 score, 0.0002 increase in MSE, and 0.0011 increase in MAE. This leads us to move onto testing a multi-variable random forest regression model.

```
Multi-Variable Random Forest Regression Model
R-squared Score: 0.3335
Mean Squared Error: 0.0291
Mean Absolute Error: 0.1239
```

Expanding to a multi-variable random forest regression model, where we incorporated the average hours spent daily on social media, daily sleeping hours, and social media addiction score in addition to the original mental health score to predict CEI regression values. We found that this regression model yielded a R^2 score of 0.3335, explaining 33.35% of the variance in CEI, with a mean squared error (MSE) of 0.0291, and a mean absolute error (MAE) of 0.1239. This model shows considerably greater regression metric scores than the previous model, with an increase of 12.31% for the R^2 score, 0.0054 decrease in MSE, and about 0.023 decrease in MAE.

4.2 Evaluation

Through our progression of regression models for this project going from linear regression to single variable and eventually to multi-variable random forest modeling, we gained insight into how young adult behaviors correlate with national cybersecurity. The similar outputs from the linear regression and single variable random forest methods indicate that mental-health scores alone produce a mostly linear relationship with cybersecurity risk, suggesting that no single behavior in youth has meaningfully indications of a complex relationship with CEI.

With our multi-variable random forest regression model having the most successful performance out of all the models we tested achieving an R^2 score of 0.3335, we can say that we have a meaningful relationship between younger generation mental health and cybersecurity exposure risk. Considering that

our project explores mental health, a subject that studies the behaviors of human subjects which is considered difficult to predict, our model is a success. The Academic Medicine & Surgery confirms this, stating “In the social sciences and psychology fields, where the behaviors of human subjects pose challenges, values as low as 0.10 to 0.30 are often considered acceptable” (Gupta 1). With our model exceeding this range, this indicates that aggregated behavioral metrics in the younger population–social media usage, sleeping patterns, social media addiction, and mental health–collectively provide a compelling signal for determining the probability of being subjected to the exposure of cybersecurity risks on a national scale.

Conclusion:

5.1 Changes After Proposal

Originally, our project was focused on analyzing individual cybersecurity behaviors of teenagers using 3 datasets related to student social media addiction, teen phone addiction, and teen cybersecurity behaviors. However, we came upon an issue specifically with our cybersecurity dataset, as we realized it was not suitable for our project because it kept showing up as having the exact same proportion for every category, which suggested that the data was synthetic rather than authentic data, despite the claims on the Kaggle site. We, therefore, couldn’t provide reliable insight into teenage online cybersecurity behavior.

As a result of our findings, we learned that our dataset would not work well with our approach, and we would need to restart our project with a replacement dataset. We replaced the synthetic dataset with a different dataset on the cybersecurity metrics of countries as a whole. We also had to remove our other phone addiction dataset, as it didn’t contain information on countries. Because of this, our approach went from “How does social media usage and mental health affect individual teen cybersecurity behaviors?” to investigating whether there is a relationship between social media and mental health patterns of young adults to national cybersecurity. We realized that social media and online presence are usually dominated by the younger population of countries, which is why we wanted to study their effect on a global scale. While there has been a change from the original proposal, our study still explores an important connection to how the younger generation’s digital use impacts the cybersecurity outcomes in the real world.

5.4 Reflection

During our project, there were several things of note that we wanted to address and think about the next time we work on a project like this. We also discussed things that we could have added or changed in order to improve our project. First, because of our initial hiccup with the changes after the proposal, it was clear to us that in order to improve this project further, we need to find more reliable datasets. In our next iteration of this project, because this is a project both of us want to improve upon even after our class, we hope to add in more reliable datasets, possibly one from a government agency that allows us to map not just students’ mental health scores but also the mental health of each country.

5.3 Final Statement

In the end, we were able to create insights on how the mental well-being of the younger population impacts the country’s cybersecurity vulnerability. Although there were some difficulties along the way, we were able to adjust and adapt to them to produce this result. There were definitely things that we would’ve done differently, but we are excited about these results, and it was an interesting topic to study and understand for both of us.

Works Cited

eMarketer. “Average Time Spent on Selected Social Media Platforms Daily among Adults in The United States in February 2024, by Age Group (in Minutes).” *Statista*, Statista Inc., 28 Mar 2024, https://www.statista.com/statistics/1484565/time-spent-social-media-us-by-age/.

Gupta, Avi, et al. “Determining a Meaningful R-Squared Value in Clinical Medicine.” Academic Medicine & Surgery, 27 Oct. 2024, academic-med-surg.scholasticahq.com/article/125154-determining-a-meaningful-r-squared-value-in-clinical-medicine, https://doi.org/10.62186/001c.125154.

Wang, Hu, et al. “Enhancing Cybersecurity Evaluation with Game Theory and MLP”. Association for Computing Machinery, New York, NY, USA, 2025. https://doi.org/10.1145/3732365.3732379.