DSA210

Screen Time VS Physical Activity: A data driven analysis

Teoman Ceyhun Etiz

33560

Supervised by Selim Balcısoy

A cell phone and a person running

AI-generated content may be incorrect.

**Introduction**

With screen time steadily rising across all age groups, especially among students, there is growing concern about its impact on physical health. This project investigates how digital screen exposure—specifically smartphone and device usage—affects an individual’s physical activity levels. The analysis is grounded in data science methods, including exploratory data analysis (EDA), hypothesis testing, and machine learning (ML) modeling.

The central aim is to determine whether increased screen time is associated with decreased physical activity and whether this behavior can be predicted or classified using modern analytical techniques.

**Data Collection**

Data was collected personally over a multi-week period using the Apple Watch SE (2nd Gen) and iPhone’s Screen Time reports. Enrichment included:

* Weather data (daily temperature, comfort index, conditions)
* Sleep data (duration)
* Academic calendar tags (e.g., exam weeks)
* Apple Health data (exercise duration, heart rate metric, etc.)

**Methodology**

After compiling the enriched dataset, I began by thoroughly cleaning and preprocessing the data to ensure consistency and accuracy. This included handling missing values, correcting data types, converting time-based fields, and standardizing numerical variables for machine learning models. The analysis followed a complete data science pipeline beginning with **Exploratory Data Analysis (EDA)**. Through various visualizations—such as scatter plots, correlation heatmaps, and time series charts—I investigated potential relationships between screen time and physical activity metrics like daily steps and exercise duration. EDA also helped identify patterns across other contextual variables such as sleep quality, weather comfort, and academic stress periods (e.g., exams).

Following EDA, I conducted **statistical hypothesis testing** to determine whether screen time had a significant effect on physical activity levels. Tests included Pearson correlation, t-tests, ANOVA, and chi-square analysis, applied across different groupings and metrics.

Finally, I implemented a **machine learning pipeline** to deepen the analysis. I applied both regression and classification models: regression to predict continuous outcomes like daily steps and exercise minutes, and classification to categorize days into high or low physical activity based on engineered thresholds. Models such as linear regression, logistic regression, and random forests were trained and evaluated using standard metrics like R², RMSE, and classification accuracy. Feature importance rankings and residual analysis were also used to interpret model behavior and validate performance.

This end-to-end approach ensured that both descriptive and predictive insights were captured, providing a robust foundation for interpreting the relationship between screen time and physical activity.

**Exploratory Data Analysis (EDA)**

The EDA phase provided a comprehensive overview of the behavioral patterns in the dataset, particularly focusing on screen time and physical activity.

A graph with a line and a line

AI-generated content may be incorrect.The distribution of daily screen time showed a slightly right-skewed pattern, with a mean of approximately 8.8 hours and a median of around 8.2 hours. This indicates that while most users spend under 10 hours on screens daily, there are outliers who spend significantly more.

A graph of a bar graph

AI-generated content may be incorrect.A breakdown of screen time by category revealed that the majority of time was allocated to **social media (average: 4.05 hours)**, followed by **entertainment (1.18 hours)**. Educational and productivity-related usage made up a minimal share, suggesting that non-essential screen time dominates daily usage.

A graph showing a comparison of a comparison between a couple of rectangular boxes

AI-generated content may be incorrect.When comparing screen time between weekdays and weekends, the data showed a modest increase on weekends, with an average of 9.23 hours compared to 8.64 hours on weekdays. This aligns with expectations that leisure time tends to increase outside of academic obligations.

A graph with a line and dots

AI-generated content may be incorrect.In examining the relationship between screen time and physical activity, scatter plots indicated a **negative correlation**between screen time and both **exercise duration** and **daily steps**. Specifically, the Pearson correlation between screen time and exercise duration was **r = -0.29** (p = 0.0858), and for screen time and steps, **r = -0.24** (p = 0.1502). While these correlations suggest that increased screen time may be associated with reduced physical activity, the relationships were **not statistically significant** at the 5% level.

A graph of a graph showing the average exercise

AI-generated content may be incorrect.Weather variables also appeared to influence activity. For instance, exercise duration was higher on **cloudy days** (51.18 minutes) than on **clear days** (35.0 minutes), possibly due to preferences for indoor workouts. Similarly, **higher temperatures** showed a slight positive correlation with exercise minutes (r = 0.30, p = 0.0805), suggesting that warmer days may encourage more physical activity.

A screenshot of a graph

AI-generated content may be incorrect.A **correlation heatmap** further supported these trends. Notably, screen time had the strongest negative correlation with **daily steps** (-0.29) and **exercise duration** (-0.26). Meanwhile, strong positive relationships were observed between **exercise duration**, **daily steps**, and **active calories burned**.

**Hypothesis Testing**

The main hypothesis tested in this study was:

**Null Hypothesis (H₀):** There is no significant relationship between screen time and physical activity.

**Alternative Hypothesis (H₁):** Screen time has a significant effect on physical activity.

1. **Pearson Correlation**
   * A **negative correlation** was observed between screen time and both exercise duration (r = -0.29) and daily steps (r = -0.24).
   * A graph with red squares and green squares

     AI-generated content may be incorrect.However, the **p-values for both correlations were greater than 0.05** (p = 0.0858 and p = 0.1502 respectively), indicating that these trends were **not statistically significant** under the 5% significance level.
2. **T-Test (High vs Low Screen Time on Exercise Duration):**
   * A t-test was conducted comparing mean exercise durations between low screen time days (≤8.25 hrs) and high screen time days (>8.25 hrs).
   * Mean exercise duration was **48.06 minutes** for low screen time and **37.94 minutes** for high screen time.
   * Despite this difference, the **p-value was 0.4956**, which is **well above the 0.05 threshold**, so the difference is **not statistically significant**.

A close-up of a white background

AI-generated content may be incorrect.

1. **One-Way ANOVA (Screen Time Tertiles vs. Exercise Duration):**
   * Screen time was divided into three groups (low, medium, high) to test for differences in average exercise duration.
   * A screenshot of a computer

     AI-generated content may be incorrect.The **F-statistic was 1.5427** with a **p-value of 0.2288**, again showing **no statistically significant differences**between the groups.
2. **Chi-Square Test (Class Load vs. Exercise Behavior):**
   * A chi-square test was performed to determine if class hours were associated with exercise levels.
   * A screenshot of a computer

     AI-generated content may be incorrect.The test yielded a **p-value of 1.0000**, indicating **no association** between class hours and physical activity.

**Conclusion:**  
Across all tests, **the null hypothesis could not be rejected**. While there is a **consistent negative trend** suggesting that increased screen time is associated with reduced physical activity, these relationships were **not statistically significant** in this dataset. This suggests that while screen time may play a role in influencing exercise behavior, its effect could be modest or obscured by other factors such as weather, class schedule, or individual habits.

**Machine Learning Analysis**

This section focuses on evaluating the ability to predict physical activity levels using screen time and other related features. Three main physical activity metrics were analyzed: exercise duration (minutes), daily step count, and an aggregated activity index. Both regression and classification models were used to assess predictive performance and identify significant variables.

Before modeling, several features were engineered to enhance predictive power. This included creating rolling averages (e.g., 3-day average screen time, 3-day average exercise duration), binary labels for high/low activity based on percentile thresholds, and aggregating physical signals like heart rate variability. Feature scaling and encoding were applied where necessary to prepare for model training.

To predict the continuous targets (exercise duration, daily steps, and activity index), two regression models were employed:

1. Linear Regression
2. Random Forest Regression

The performance of the models was assessed using the coefficient of determination (R²) and root mean squared error (RMSE).

For exercise duration prediction, the Linear Regression model achieved an R² of 0.383 and RMSE of 30.97. The prediction vs. actual plot revealed a general alignment but with notable residual variance, especially for low or high exercise values. The Random Forest model showed a much stronger performance with an R² of 0.819 and RMSE of 16.79, and its residuals were more symmetrically distributed. This indicates that Random Forest captured non-linear patterns more effectively.

A graph with red lines

AI-generated content may be incorrect.A graph with a red line

AI-generated content may be incorrect.

A graph with red lines and numbers

AI-generated content may be incorrect.A graph with a red line

AI-generated content may be incorrect.

A graph with a red line

AI-generated content may be incorrect.A graph with red lines and a red line

AI-generated content may be incorrect.In the case of daily step count, the Linear Regression model failed to fit the data adequately, yielding a negative R² of -0.01 and RMSE of 5353.47. The prediction scatter plot and residual histogram showed random scatter, suggesting poor model fit. However, Random Forest Regression improved the prediction accuracy considerably, with an R² of 0.410 and RMSE of 4089.74.

A graph with red lines and numbers

AI-generated content may be incorrect.A graph with a red line

AI-generated content may be incorrect.

For the activity index, which is a normalized metric combining multiple physical activity indicators, the Linear Regression model performed modestly (R² = 0.392, RMSE = 0.19), while the Random Forest model again performed substantially better with R² = 0.749 and RMSE = 0.12. Visualizations for residuals and prediction alignment confirmed that the Random Forest model captured the pattern more accurately.

A graph with a red line

AI-generated content may be incorrect.

A graph with red lines and black text

AI-generated content may be incorrect.

A graph with a red line and blue dots

AI-generated content may be incorrect.

A graph with red lines and a red line

AI-generated content may be incorrect.

In addition to regression tasks, classification models were used to predict binary outcomes such as whether a day would be a high exercise day or a high step day.

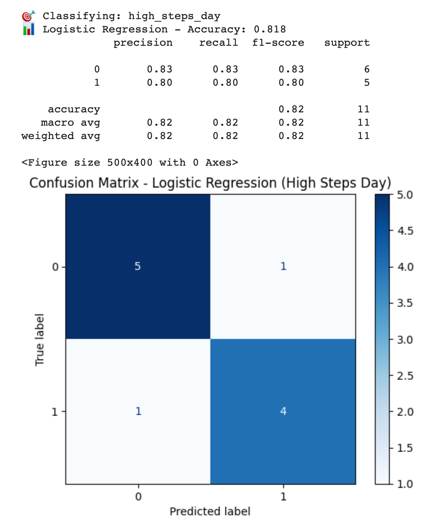
Methods used:

1. Logistic Regression
2. Random Forest Classification

The results were evaluated using accuracy, precision, recall, F1-score, confusion matrices, and ROC-AUC scores.

A graph of a logistic regression

AI-generated content may be incorrect.A screenshot of a graph

AI-generated content may be incorrect.A screenshot of a computer screen

AI-generated content may be incorrect.A screenshot of a graph

AI-generated content may be incorrect.For high exercise days, Logistic Regression achieved an accuracy of **81.8%** with an **AUC of 0.97**, indicating strong sensitivity and specificity. However, the Random Forest Classifier significantly outperformed it, achieving **100% accuracy** and an **AUC of 1.0**, suggesting perfect classification on the test set. While impressive, such perfect performance should be interpreted cautiously — it may reflect underlying data separability, or potential overfitting if the dataset is small or imbalanced.

A graph of a curve

AI-generated content may be incorrect.A graph of a graph

AI-generated content may be incorrect.A graph of a logistic regression

AI-generated content may be incorrect.

For high step days, both models performed similarly in terms of accuracy (**81.8%**), though the Random Forest still achieved a slightly higher AUC (**0.93**) than Logistic Regression (**0.87**), reflecting better overall discriminative ability.

These results indicate that physical activity classification is feasible with the current feature set, particularly when using tree-based models that capture non-linear interactions. Features such as **heart rate variability**, **screen time averages**, and **past activity history** likely played key roles in the classification decision, as revealed by feature importance analysis.

A feature importance analysis using the Random Forest model revealed that the most predictive feature for activity levels was heart rate difference, followed by average daily heart rate and recent exercise history (3-day average). Environmental features such as humidity and screen time (3-day average) also A graph with blue bars

AI-generated content may be incorrect.contributed, though to a lesser extent.

These findings indicate that Random Forest models are superior to linear approaches in capturing the complex, non-linear relationships between screen time and physical activity. While screen time showed some predictive value, physiological and contextual factors (e.g., heart rate metrics, humidity) played a larger role in predicting exercise behavior. This emphasizes the importance of considering a broader set of features beyond screen time alone.

**Final Findings**

The goal of this project was to understand whether screen time affects physical activity levels. The analysis combined three key approaches: exploratory data analysis (EDA), statistical hypothesis testing, and predictive modeling through machine learning.

From the EDA, a negative trend was observed—days with higher screen time generally corresponded to lower levels of physical activity (fewer steps and shorter exercise duration). However, this relationship was not very strong visually and showed a lot of variability.

Statistical tests supported this initial observation but did not confirm it with strong significance. The Pearson correlation showed a weak to moderate negative relationship between screen time and exercise minutes, but the p-value was above the 0.05 threshold. Similarly, the t-test and ANOVA comparing physical activity across high and low screen time days did not find statistically significant differences. Therefore, we could not reject the null hypothesis that screen time has no effect on physical activity.

Machine learning models, particularly the Random Forest regressors and classifiers, provided more nuanced results. These models were able to predict physical activity metrics like exercise duration, step count, and activity index with moderate to high accuracy. Screen time—especially when averaged over multiple days—contributed to the model’s predictions, but it was not the most important feature. Physiological indicators (like heart rate difference and average heart rate) and environmental factors (like humidity) were often stronger predictors.

In conclusion, screen time shows a weak but consistent negative relationship with physical activity. While it plays a role, it is not the dominant factor affecting physical activity levels. More accurate predictions come from combining screen time with physiological and contextual data. This means that screen time alone is not a reliable indicator of physical activity but can be useful when considered alongside other variables.