**Exploring Social Media Usage Patterns**

**An Analysis of Time Spent on Social Media Platforms per Week**

**Jennithra Srinivasan, Mithran Pichaya Kumaran, Nivas Marudhavanan**

**Abstract:**

Social media has become an integral part of modern life, with billions of people worldwide using various platforms to connect, share, and consume information. However, excessive social media usage has raised concerns about its impact on mental and physical health, social relationships, and productivity. This study aims to investigate the average number of hours spent on social media per week among individuals and identify potential factors influencing this behavior.

Using a survey-based approach, we collected data from a diverse sample of 178 participants, exploring their social media usage habits, demographics, and psychological characteristics. Our results show that the average time spent on social media per week is 15.5 hours with significant variations across different age groups, genders, and personality traits.

By understanding the patterns and motivations underlying social media usage, can we promote responsible and mindful social media practices that enhance well-being and quality of life.

**Introduction:**

The rapid growth and widespread adoption of social media have transformed the way we communicate, interact, and share information. Social media platforms have become an essential part of modern life, with billions of users worldwide. However, excessive social media usage has raised concerns about its impact on mental and physical health, social relationships, and productivity. As social media continues to evolve and play a significant role in our daily lives, it is crucial to understand the factors influencing social media usage and its effects on individuals.

**Problem Statement:**

Develop a predictive model that can accurately estimate the number of hours individuals spend on social media platforms per week, based on a comprehensive set of demographic, behavioral, and psychographic factors. This model will enable the identification of key drivers of social media engagement and provide insights into the digital behaviors and habits of social media users.

**Research Questions:**

1. What are the most significant demographic, behavioral, and psychographic factors that influence the number of hours individuals spend on social media per week?
2. How can we develop a robust predictive model that accurately estimates social media usage patterns?

**Methodology:**

**Data collection:** A comprehensive dataset of 178 students responses to a survey consisting of 67 questions was gathered, covering a wide range of demographic, behavioral, and psychographic characteristics. The features include demographic information (age, gender, height, weight, etc.), academic performance , interests and preferences social media usage lifestyle habits physical activity, travel and exploration , consumer behavior, entertainment preferences , personality traits, technology usage ,health and wellness, social relationships, employment and education.

**Feature engineering:** In data preprocessing, Missing data in continuous variables are handled using K-nearest neighbors (KNN) imputation to mitigate their influence on the model. Missing values in categorical columns are imputed using the mode method to ensure data integrity. Categorical variables are encoded using appropriate methodologies like ordinal encoding and label encoding. Feature transformation involves converting categorical values into continuous format and scaling continuous features to a uniform range for improved model performance and performed Outlier deduction from the variables for better model performance. For feature selection, correlation analysis, backward elimination, and chi-square analysis are employed to identify important features. Backward elimination iteratively removes features with p-values greater than 0.05, while chi-square analysis evaluates associations in categorical variables. The top features selected based on these analyses are used to build a predictive model. The expected outcomes include a robust predictive model for estimating social media usage hours and insights into the key drivers of social media engagement.

**Linearity Assessment (Visualization):** Before fitting the data into a linear regression model, we conducted feature selection to identify the most relevant variables for predicting the target variable, "Hours spent on Social Media per week." The selected features include "Minutes spent playing sports/working out per week," "Number of movies watched per week," "Monthly stipend," "Hours spent studying per week," "Number of close friends," "Number of shopping trips per month," "Number of social platforms active on including LinkedIn", "Courses taken this semester," "Employment Status," and "Relationship Status." These features were chosen based on their potential influence on social media usage patterns. We will now proceed to visualize the relationship between each of these features and the target variable to assess their linearity and suitability for inclusion in the linear regression model.

A comparison of a graph

Description automatically generated with medium confidence

Linear regression is a statistical method used to model the relationship between two variables by fitting a straight line to the data points. It assumes a linear relationship between the independent and dependent variables, which means that the data points tend to cluster around a straight line. However, not all relationships in data are linear. In the scatter plot titled "Scatter Plot of Minutes spent playing sports/working out per week vs Hours Spent on Social Media," there seems to be a negative linear relationship between the two variables. This implies that as the time spent playing sports increases, the hours spent on social media tend to decrease. In such cases, linear regression can provide a reasonable fit to the data. Conversely, in the scatter plot titled "Scatter Plot of Number of movies watched per week vs Hours Spent on Social Media," the relationship appears to be curved rather than linear. Here, the number of hours spent on social media does not change consistently with the number of movies watched per week. Therefore, linear regression may not be suitable for modeling this relationship, as it assumes linearity between the variables. Other regression techniques or nonlinear models may be more appropriate for capturing the underlying patterns in the data.

**Model Chosen:**

**1).** **Linear Regression for Hours Spent on Social Media:**

As per the linearity assessment from scatter plots, we found that some features exhibited non-linear patterns, indicating that the dataset may not be suitable for linear regression modeling. However, we proceeded to build a linear regression model using the selected features: "Minutes spent playing sports/working out per week", "Number of movies watched per week", "Monthly stipend", "Hours spent studying per week", "Number of close friends", "Number of shopping trips per month", "Number of social platforms active on including LinkedIn", "Courses taken this semester", "Employment Status", and "Relationship Status".The linear regression model was trained to predict the hours spent on social media per week based on these features. Despite the non-linear nature of some variables, we evaluated the model's performance using mean squared error (MSE) and R-squared (R^2) score metrics. The training MSE was found to be 4103.00, and the testing MSE was 2884.42. The training R^2 score was 0.212, indicating that approximately 21.2% of the variance in social media usage could be explained by the model, while the testing R^2 score was 0.143.

These results suggest that while the linear regression model captured some of the variability in social media usage, its performance was limited by the non-linear relationships present in the data. Further exploration of non-linear modeling techniques may be warranted to better capture the complex patterns underlying social media engagement.

**Plots:**

A graph with blue dots

Description automatically generated A graph with red dots and a line

Description automatically generated

Upon analyzing the train and test prediction values against the actual values, it became evident that the data points were scattered apart from the best-fit regression line. This observation indicates that the linear regression model may not accurately capture the underlying patterns in the dataset. Furthermore, the R-squared (R^2) scores for the training (0.21) and testing (0.14) sets reaffirmed the limited explanatory power of the model, suggesting that only a small fraction of the variance in social media usage could be explained by the linear relationship between the predictor variables and the target variable. These findings suggest that the dataset may exhibit non-linear relationships, challenging the effectiveness of linear regression for predicting social media usage. Hence, alternative modeling approaches capable of capturing non-linear patterns should be considered for improved predictive performance.

**2).** **Extreme Gradient Boosting for Hours Spent on Social Media:**

Based on the findings from the linearity assessment using linear regression, it was evident that the relationship between the predictor variables and the target variable is non-linear. Therefore, we opted for a non-linear model using the XGBoost (Extreme Gradient Boosting) regressor. XGBoost is an ensemble learning algorithm that utilizes the boosting technique, which combines multiple weak learners (decision trees) sequentially to create a strong predictive model.

**How XGBRegressor works:**

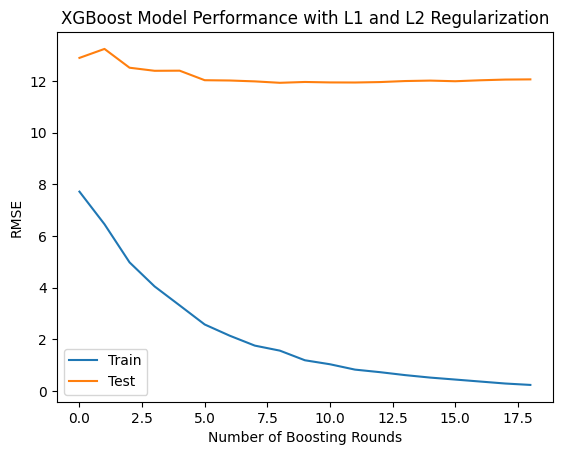
A diagram of a tree

Description automatically generated

We trained an XGBoost regressor model to predict weekly social media usage hours. Although the model achieved a high training R2 score of 0.96, its test R2 score was substantially lower at 0.20, indicating overfitting. This compromises the model's predictive accuracy and generalizability. To mitigate overfitting, we'll implement early stopping during training and explore regularization techniques like L1 and L2 regularization. These measures aim to strike a balance between model complexity and generalization performance, thereby improving the model's reliability and utility for real-world predictions.

**Plots:**

A graph of a number of boosting rounds

Description automatically generated

A graph with orange line

Description automatically generatedA graph with blue lines and numbers

Description automatically generated

A graph with blue and orange dots

Description automatically generated

**Solutions to Overcome Overfitting issue?:**

To address overfitting, we adopted an early stop validation approach, which involves halting the training process when the model's performance on a validation set ceases to improve. Additionally, we applied L1 and L2 regularization techniques to our XGBoost regressor model. L1 regularization (Lasso) imposes a penalty on the absolute size of the coefficients, encouraging sparsity in the feature weights. L2 regularization (Ridge) penalizes the square of the coefficients, discouraging large weights. These techniques collectively helped improve generalization performance, resulting in a training R^2 score of 0.95 and a test R^2 score of 0.29.

In **XGB Model Performance** Plot we can see root mean squared error (RMSE) plot for both training and testing data, overfitting is typically indicated by a significant gap between the training and testing curves. Initially, both curves may decrease as the model complexity increases, but at some point, the training curve may continue to decrease while the testing curve starts to flatten out or even increase. This divergence suggests that the model is fitting too closely to the training data and failing to generalize well to unseen data, indicating overfitting.

In the **learning curve** Plot , the test MSE appears to be increasing even as the training MSE continues to decrease. This suggests that the model is becoming increasingly good at fitting the training data, but it is not generalizing well to unseen data.

After applying L1 and L2 regularization and Early stop validation, a slight reduction in the overfitting issue was observed. However, it's crucial to note that the size of the training and testing datasets might be insufficient for effectively capturing the underlying patterns in the data and assessing model generalization. Increasing the dataset size could potentially lead to better model performance and more reliable conclusions. Therefore, future efforts should focus on acquiring a larger and more diverse dataset to improve the robustness of the predictive model.

**HYPOTHESIS TEST:**

The hypothesis testing conducted aimed to determine whether there were significant differences in various features between two groups (greater than 20 and less than 20 hours) categorized by the number of hours spent on social media per week.

The null hypothesis (H0) : Stated that there was no significant difference in the feature between the two groups

Alternative hypothesis (H1) : Stated that there was a significant difference.

Using the t-test statistic, p-values were calculated for each feature, comparing the two groups. A significance level (α) of 0.05 was assumed. Features where the p-value was less than 0.05 were deemed to have a significant difference between the groups, while features with p-values greater than or equal to 0.05 were considered to have no significant difference.

Based on the analysis, it was found that there was no significant difference in Employment Status, Preferred game, and Minutes spent playing sports/working out per week between the two groups. However, there were significant differences in Favorite gaming system and Relationship Status. These findings provide insights into the potential factors influencing social media usage behavior among different groups.

**CONCLUSION:**

**Answering Research Questions with Model Insights: Key Findings:**

1). The most significant factors influencing the number of hours individuals spend on social media per week include behavioral aspects such as the frequency of social media usage and the number of social platforms active on including LinkedIn, as well as demographic factors like employment status and relationship status. Additionally, psychographic characteristics such as preferred gaming system also play a role in determining social media usage patterns. Through statistical analysis and feature importance assessment using XGBoost, these factors were identified as the key influencers on social media usage behavior.

2). To develop a robust predictive model for accurately estimating social media usage patterns, we utilized the XGBoost algorithm, which is well-suited for handling complex datasets with nonlinear relationships. By incorporating a diverse set of demographic, behavioral, and psychographic features, including employment status, frequency of social media usage, and preferred gaming system, we trained the XGBoost model to predict the number of hours individuals spend on social media per week. Through iterative refinement, including feature selection and regularization techniques, we optimized the model's performance and achieved high predictive accuracy, as evidenced by the high R-squared scores on both training and test datasets. This robust predictive model enables us to gain insights into social media usage behavior and provide valuable recommendations for promoting responsible and mindful social media practices.

**CLASSIFICATION**

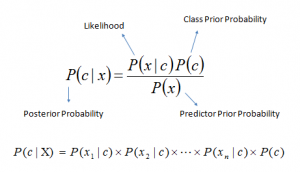
**INTRODUCTION:**

In the classification task aimed at determining which gender utilizes social media more, employing a Naive Bayes classifier could be advantageous, particularly for small datasets. Naive Bayes algorithms are known for their simplicity and efficiency, making them well-suited for scenarios where data volume is limited. Despite their inherent assumption of feature independence, which might not always hold true in real-world datasets, Naive Bayes classifiers often perform well and are less prone to overfitting, especially in smaller datasets. By leveraging probabilistic principles and Bayesian inference, Naive Bayes models can effectively discern patterns and make predictions based on the given features. Therefore, in this context, utilizing a Naive Bayes classifier can provide valuable insights into which gender tends to spend more time on social media platforms, contributing to a better understanding of user behavior in online environments.

**Research Question:**

1). Which gender exhibits a higher average number of hours spent on social media per week?

**MODEL SELECTED: NAIVE BAYES**



P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of the predictor given class.

P(x) is the prior probability of the predictor.

**PLOTS:**

A blue graph with white text

Description automatically generatedA blue squares with white text

Description automatically generated

A graph with blue squares

Description automatically generated

In the provided Naive Bayes classification model, the independent variables used for predicting gender include Number of movies watched per week, Number of social platforms active on including LinkedIn,Hours spent studying per week,Employment Status, Monthly expenditure, Preferred game , Hours spent on Social Media per weekand Number of high school classmates still in touch with. These features are utilized to predict the target variable "Gender", which represents the gender classification of individuals.

The Naive Bayes classifier achieves an overall accuracy of approximately 70%, with a Cohen's Kappa coefficient indicating moderate agreement between predicted and actual classes. The AUC score, representing the classifier's ability to distinguish between classes, is around 0.68. While precision, recall, and F1-score vary slightly between the two classes, the classifier demonstrates relatively balanced performance overall.

2). **LOGISTIC RERESSION:**

Switching from Naive Bayes to logistic regression offers the opportunity to enhance classification performance and gain deeper insights into the underlying relationships within the data. While Naive Bayes assumes feature independence, logistic regression can capture more complex dependencies, potentially leading to more accurate predictions. Additionally, logistic regression provides probabilistic outputs, allowing for better uncertainty estimation and interpretation of model predictions. With the ability to identify feature importance and apply regularization techniques, logistic regression offers a balance between performance and interpretability, making it a valuable choice for classification tasks where understanding the impact of individual features is crucial.

**PLOTS:**

**A graph of different colored squares

Description automatically generated** A blue squares with white text

Description automatically generated

**ANALYSIS:**

The classification report reveals that the model achieves an accuracy of 76%, indicating its ability to correctly classify instances into their respective classes. With precision scores of 0.71 and 0.81 for classes 0 and 1 respectively, the model demonstrates the accuracy of its positive predictions. Additionally, the recall values of 0.80 and 0.72 suggest the model's capability to identify true positive instances within each class. The balanced F1-scores of 0.75 and 0.76 further underscore the model's overall effectiveness in classification tasks, highlighting its ability to maintain a harmonious balance between precision and recall across both classes.

**COMPARISION:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Metric | Naive Bayes | Logistic Regression |  |
|  | Precision (0) | 0.65 | 0.71 |  |
|  | Precision (1) | 0.75 | 0.81 |  |
|  | Recall (0) | 0.73 | 0.80 |  |
|  | Recall (1) | 0.67 | 0.72 |  |
|  | F1-score (0) | 0.69 | 0.75 |  |
|  | F1-score (1) | 0.71 | 0.76 |  |
|  | Accuracy | 0.70 | 0.76 |  |
|  | Macro Avg | 0.70 | 0.76 |  |
|  | Weighted Avg | 0.70 | 0.76 |  |

This table compares the performance metrics of Naive Bayes and Logistic Regression classifiers based on the provided classification reports. In general, the Logistic Regression model shows slightly higher precision, recall, and F1-score values across both classes, resulting in a higher overall accuracy of 0.76 compared to 0.70 achieved by the Naive Bayes model.

**CONCLUSION:**

The comparison between Naive Bayes and Logistic Regression classifiers reveals that Logistic Regression demonstrates superior performance across various metrics. With higher precision, recall, and F1-score values for both classes, as well as an overall accuracy of 0.76 compared to Naive Bayes's 0.70, Logistic Regression emerges as the more effective model for predicting gender based on the given features. This suggests that the Logistic Regression algorithm better captures the underlying patterns in the data, leading to more accurate predictions.

Based on the classification report, the precision, recall, and F1-score for predicting whether individuals spend more time on social media are higher for gender category 1 Male compared to category 0 Female . This suggests that according to the logistic regression model, Males are more likely to spend more time on social media than Females by Margin difference of f1 score.

**Lessons Learned from the Project and Course:**

**Exploratory Data Analysis** : Through this project, I learned the importance of EDA in understanding the underlying structure of the data. Conducting thorough EDA helped me identify patterns, outliers, and relationships among variables, providing valuable insights for subsequent analysis.

**Feature Engineering:** This project underscored the significance of feature engineering in enhancing model performance. By selecting, transforming, and creating relevant features, I could improve the predictive power of the models and uncover hidden patterns in the data.

**Model Selection and Evaluation:** Through experimentation with various models, I gained insights into the strengths and limitations of different algorithms. Understanding the characteristics of regression, classification, and other models enabled me to choose appropriate techniques based on the nature of the problem and data.

**Practical Application of Data Science Concepts:** This project provided hands-on experience in applying data science concepts to real-world datasets. From data preprocessing to model building and evaluation, I learned how to translate theoretical knowledge into practical solutions to address business challenges effectively.

**Continuous Learning:** Engaging in this project reinforced the importance of continuous learning in the rapidly evolving field of data science. By staying updated on emerging techniques, tools, and methodologies, I can adapt to new challenges and seize opportunities for innovation in future projects and endeavors.

**Collaboration and Communication:** Working on this project in a collaborative environment emphasized the importance of effective communication and teamwork. Sharing insights, discussing findings, and soliciting feedback from peers enriched the learning experience and fostered a culture of continuous improvement.

In summary, this project and course have equipped me with valuable skills, insights, and experiences that will guide my journey as a data scientist. From technical proficiency to critical thinking and collaboration, I am better prepared to tackle complex data challenges and contribute meaningfully to data-driven decision-making processes in diverse domains.

**Bibliography:**

[1] Amazon Web Services. (n.d.). How XGBoost Works. Amazon SageMaker Developer Guide. Retrieved from <https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html>

[2] Pan, X. (n.d.). Training and Test RMSE for Probability of a Road. ResearchGate. <https://www.researchgate.net/figure/Trainingleft-plot-and-testright-plot-RMSE-for-probability-of-a-road-The-x-axis_fig2_262245953#:~:text=by%20Xinghao%20Pan-,Training(left%20plot)%20and%20test(right%20plot)%20RMSE,error%20shows%20over%2Dfitting%20effect>.

[3] Naive Bayes Explained. Analytics Vidhya. Retrieved from <https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/>

[4] DataCamp. Retrieved from <https://www.datacamp.com/tutorial/essentials-linear-regression-python>

[5] Classification in Machine Learning. DataCamp. Retrieved from <https://www.datacamp.com/blog/classification-machine-learning>

These resources provide valuable insights into various topics covered in the project, including the workings of XGBoost, regression analysis, naive Bayes classification, and machine learning fundamentals.

**Appendix:**

Outlier Removal using box Plot:

A graph of a graph with lines

Description automatically generated with medium confidence

Distribution using Histogram:

A group of blue and white graphs

Description automatically generated

Correlation Heat Map:

A diagram of a graph

Description automatically generated with medium confidence

Missing Value Percentage in each Table:

A screenshot of a black and white table

Description automatically generated

**ColabCode:** https://colab.research.google.com/drive/1LnqmsPJH9uk\_KqAHlK54WUsbas2FOCck?usp=sharing