International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 13 Issue 3 ǁ March 2025 ǁ PP. 85-87

Social Media Addiction Prediction Using Machine Learning

S.Arockiya Gresa, R.Janani, B.Kavitha, Dr. P. Meenakshi Sundari, MSc., M.Phil., SET, Ph.D.

Department of Computer Science, Kamaraj University, Fatima College, Madurai

Date of Submission: 09-03-2025 Date of acceptance: 23-03-2025

I. INTRODUCTION

From businesses to marketers to influencers, social media platforms create vast amounts of data that can be used to predict user behavior, engagement, and content trends. These predictions are made using machine learning, specifically logistic regression when it comes to the classification of the events as a task involving just two target outcomes. Logistic regression examines the relationships between variables such as the above, giving us insights into what actually drives your social media success.

Machine learnings lets you predict engagement behaviours (sentiment, purchase intent, etc), optimize your marketing strategy, and even predict whether or not something will go viral. With the help of machine learning and logistic regression, businesses can improve targeting, strengthen content strategies, and make better business decisions for effective social media management.

DATA COLLECTION

For this study, data was collected from people through carefully designed questions in surveys and questionnaires. These questions aimed to gather insights into the user's social media usage, engagement behaviour, and psychological effects. The goal was to create a dataset with sufficient information to predict social media addiction

a. Survey Design

The survey includes quantitative and qualitative questions regarding social media use and its emotional and psychological effects. The questions were broken down into a few major categories:

1. Demographic Information:
Age: What is your age?
Gender: What is your gender?

2. Social Media Usage Patterns:
How many hours per day do you spend on social media?
How often do you check social media notifications?
Do you use social media while eating meals?

3. Emotional and Psychological Impact:
Do you use social media as an escape from stress, sadness, or boredom?
Do you feel that social media affects your real-life social interactions?
Do you compare your life to others on social media and feel unhappy?

4. Engagement Behaviour:
Do you experience eye strain, headaches, or fatigue due to prolonged social media use?
How often do you use social media before going to sleep?
How often do you check social media notifications?

DATA PREPROCESSING

Preparing the dataset for analysis, Data preprocessing is a key step. The raw data may also be noised, contains missingness, duplicates and outliers which has to be dealt with before modelling.

HANDLING MISSING DATA

It is a common problem with real world datasets that some data is missing. In this case:

- **Categorical Features:** For categorical features with missing values (i.e., "Gender", "Occupation" etc.), we imputed the most frequent value (mode imputation).
- **Numerical Features**: For the numerical features that are missing values (the time spent on Social-Media, Psychological Scores), we applied mean imputation for features that contain less propagated missing data, and a median imputation approach was applied on features containing a higher percentage of missing data.

.

Outlier Detection and Handling

Outliers have a very influential role on model performance and having such high end values really skew the results if you are using algorithms like linear regression. We used the following techniques to detect and treat outliers:

- **Boxplots**: Boxplots were used to visually check features for outliers.
- **Z-Score Method**: Normalizing the distribution for numerical features we calculated the Z-score and dropped values for Z-score > 3 (far away from mean).

Encoding Categorical Variables

Categorical features (like "Gender" or "Occupation") were encoded as follows:

- Nominal Variables (e.g. "Gender"): For nominal variables, we used one-hot encoding, which creates binary columns for each category.
- Encoding featuresUser defined sequential labels for the encode operation. Ordinal Encoding: We used ordinal sorting for "Occupation" and hence applied OrdinalEncoder (for sequential categorical variables).

Feature Scaling

Some machine learning algorithms require scaled features, such as Support Vector Machines SVM and k-Nearest Neighbors k-NN. We applied standardization, or z-score normalization, to ensure that all numerical features had a mean of 0 and a standard deviation of 1. This was specifically applicable to the features "Time Spent on Social Media" and "Age."

Feature Selection

After deep consideration, we have outlined the key features that help enhance the model's effectiveness as well as overfitting prevention measures.

- **Correlation Analysis:** In particular, we used the correlation matrix to note the highly correlated features. Those features with high correlations (above 0.9) were filtered out in order to reduce multicellularity
- **Recursive Feature Elimination (RFE):** We applied RFE in a recursive manner such that it eliminated the less important features and thus ensured that the model concentrated on the most influential ones only.

Data Analysis

Exploratory Data Analysis (EDA)

EDA helps in understanding the underlying structure of the data and uncovering patterns. We used various statistical and visualization techniques to analyse the data:

- **Time Spent on social media**: A histogram showed a positive skew, indicating that most users spent a moderate amount of time on social media, but a few users spent excessive hours.
- **Age vs. Addiction:** A box plot showed that younger individuals, particularly those in the age group of 16-30, were more likely to exhibit addiction behaviour. This suggests that age is a significant factor in predicting social media addiction.
- **Psychological Factors**: We found a significant positive correlation between social media addiction and higher levels of anxiety and depression. Scatter plots revealed that users with higher anxiety scores tended to be more addicted to social media.
- **Gender and Addiction**: The distribution of addiction between genders showed no significant difference, suggesting that both males and females were equally at risk for addiction based on the features in the dataset.

Correlation Analysis

A heatmap was used to identify relationships between numerical variables and the target variable (social media addiction):

- Strong correlations were found between "Time Spent on Social Media" and "Addiction Risk" (0.76), implying that longer social media usage is a key predictor of addiction.
- Moderate correlations were observed between "Psychological Factors" (anxiety, depression) and addiction risk (0.63 and 0.68, respectively), highlighting the psychological aspect of addiction.

Feature Importance

Using techniques like Random Forest and XGBoost, we evaluated the importance of each feature in predicting addiction:

- **Time Spent on social media** was the most important feature.
- **Anxiety and Depression Scores** also emerged as key predictors.
- **Age** and **Frequency of Use** were also found to be significant, with younger users and frequent social media users showing higher addiction risk.

II. Conclusion

This report highlights the key measures taken to preprocess and analyse data for predicting social media addiction. Some of the critical findings are as follows:

• Two important predictors of social media addiction are the time spent on social media, and psychological factors like anxiety and depression.

• Addiction is more prevalent among younger individuals, but gender and occupation have lesser effects on addictivebehavior.

• Data that has been pre-processed, including missing value treatments, one-hot encoding of categorical variables and normalizing features, is what is required prior to training the machine learning model. This data and framework, further refined, can support the development of predictive models aimed at identifying individuals susceptible to social media addiction, thereby enabling early interventions for healthier digital behavior.

OUTPUT

```
Model Accuracy: 0.93

Enter user details:
What is your age?
1. under 25
2. 25-35
3. 35-55
Enter your choice (number): 3
What is your gender?
1. Male
2. Female
3. Other
Enter your choice (number): 2
How much time do you spend on social media daily?
1. Less than 1 hour
2. 1-3 hours
3. 3-5 hours
4. More than 5 hours
Enter your choice (number): 4
How often do you check social media notifications?
1. A few times a day
2. Every hour
3. Every 15-30 minutes
4. Constantly
Enter your choice (number): 4
Do you use social media while eating meals?
1. Never
2. Rarely
3. Sometimes
4. Almost always
Enter your choice (number): 3
How often do you use social media before going to sleep?
1. Never
2. Occasionally
3. Frequently
4. Always
Enter your choice (number): 2

Do you experience eye strain, headaches, or fatigue due to prolonged social media use?
1. Never
2. Rarely
3. Sometimes
4. Frequently
Enter your choice (number): 4

Prediction: The person is not addicted to social media.

☑ Dataset updated successfully in 'social media.csv' with original user inputs.
```