DATATHON Team overview

Team Name: BE2_CE

Problem statement no and name: 2

Understanding and sight of data (under 10 points):

1. The dataset contains information about social media usage patterns, engagement levels, user demographics, and dominant emotions.

2. Extensive data cleaning was carried out using manual and automated scripts efforts. Duplicated feature values are mentioned which convey the same meaning (Eg. Overweight/Obese etc.)

3. **Feature Engineering**: Created new features such as *Total_Engagement* (sum of likes, comments, and messages) and *Engagement_Received* to analyze user interactions better.

4. Sleep Dataset exhibit uniform activity patterns (Eg. Doctors record 8000 steps and 70BPM heart rate in all records.
There are null values in numerical columns which need to be dealt with accordingly.

5.Social Media data is between 18-30. Hence to connect both the data conclusions, we need to work in 18-30 age category

6. Used **SMOTE (Synthetic Minority Over-sampling Technique)** to balance class distribution, particularly for underrepresented age groups and emotions.

7. **Exploratory Data Analysis (EDA)**:

- Visualized the distribution of engagement across different emotions.
- Used **box plots** and **count plots** to analyze patterns in user behavior.
- Examined relationships between social media activity and engagement.

8. **Performance Evaluation**:

- Models were evaluated using accuracy, **F1-score (weighted)**, and other relevant metrics.
- Conducted train-test splits to validate performance effectively.

9. **Deployment & Usability**:

- The best-performing model can be used for engagement prediction. This can be used by medical and health institutes for predicting the mental and physical health of the individual based on his social media and sleep profile.

- Data insights can help social media platforms optimize content strategies based on user emotions.

Cleaning and data transformation:

Description	Y/N	Method used
Skewed data check	Y	.skew() and checking the normal curve
Duplicity check	Y	Created Automated python scripts
Null Check	Y	.isna().sum() and checking the overall values
Data type check	Y	.dtypes()
Range check	Y	Used .describe() method to understand statistics
Feature importance check	Y	Correlation matrix to find importance with respect to one other
Sample imputer used	Y	Scaled date only for sleep dataset
Feature engineering used	Y	Created new features by reducing dimensions
Standard deviation check	Y	As part of describe() feature
Variance check	Y	As part of describe() feature
Bias check	N	

Data Visualisation:

Description	Y/N	Number and feature used
Corelation Matrix	Y	Used for both sleep and socialmedia analysis. Total 10
Pair plots	N	
Data definition check	Y	Box plot and various visualization techniques

Model selection

Description	Since data is extremely small and repetitive, we used Random Forest and XGBoost
No of model tested	5
Model select criteria	Accuracy ability to handle low data

DATATHON Team overview

Method used to select models	Feature engineering and feature importnce
Model accuracy	12%
Confusion matrix	Yes done
R^2 Value	0.82

Challenges:

Challenges	Solution
• Analyze user behavior across multiple social media platforms (YouTube, Instagram, Facebook, WhatsApp, etc.) based on various entity like age groups.	Used one hot encoding and label encoder for category variables as well as for the target variable. Random forest model was used for finding out the platform based on various other input features.
• Develop a **predictive model** that forecasts which social media platform a person is likely to use based on their age and engagement patterns.	Clustering the data groups on basis of Platform. However, inspite of normalization and other techniques, overfitting couldn't be reduced. A random forest and decision tree method was used for the same.
• Create an **interactive data visualization dashboard** to present insights in an intuitive and engaging way.	Different visualization methods were used like correlation matrix,heatmap,Axes3D boxplot analysis and other interactive graphs.
• Identify and predict **potential mental and physical health effects** associated with different usage patterns.	2 different models were developed. One for the physical health part and one for mental health part. (Similar to ensemble part); where in we take predictions from each part and then combine the outcomes to get the final results.