Team Name: CodeCrew

Problem statement no and name:2

Understanding and sight of data (under 10 points):

1. Datasets Overview: All datasets have a common identifier: ‘User\_ID’ which is useful for merging.
2. **User Demographics:** Columns include Age, Gender, and Platform. Gender is categorical (Male, Female), Age is numeric. Some missing or NaN values in demographic data
3. **Social Media Behavior:** Features like Daily\_Usage\_Time, Posts\_Per\_Day, Likes\_Received\_Per\_Day, Comments\_Received\_Per\_Day, Messages\_Sent\_Per\_Day. These reflect how active users are on social platforms. Some columns show skewness (e.g., heavy users with extreme values)
4. Emotional Insights: Column- Dominant\_Emotion (e.g., Happiness, Sadness, etc.) Indicates psychological/emotional state derived from usage or engagement.
5. Sleep Dataset Overview: Features: Sleep\_Duration, Sleep\_Quality, Stress\_Level, Activity\_Level, etc. Key for analyzing mental and physical well-being in relation to social media use.
6. Merge Potential: All datasets include User\_ID → can be joined for holistic analysis across behavior + health.
7. Data Cleaning Needed: Some rows are **fully NaN** and must be dropped.Columns like Unnamed: 10 are irrelevant and should be removed. Some numeric columns are stored as **object type** and must be converted.
8. Missing Values & Duplicates: Several missing values across datasets — especially in test/val. Needs imputation (e.g., mean for numerical, mode for categorical). Few duplicate rows or duplicate User\_IDs.
9. Feature Distributions: Variables like Daily\_Usage\_Time, Posts\_Per\_Day are **right-skewed**. Use histograms, box plots, and log transformations to handle skewness/outliers.
10. Correlation & Modeling Potential: Strong correlations expected between usage and engagement metrics. High potential for: **Predictive modeling** (e.g., predict stress/sleep/emotion). **Clustering** (user segmentation).

Cleaning and data transformation:

|  |  |  |
| --- | --- | --- |
| **Description** | **Y/N** | **Method used** |
| Skewed data check | y | skew() function used in statistical analysis |
| Duplicity check | y | duplicated().sum() for duplicate detection |
| Null Check | y | isnull().sum() for missing value detection |
| Data type check | y | dtypes attribute used for verification |
| Range check | y | min()/max() used for value range validation |
| Feature importance check | y | RandomForest feature\_importances\_ used |
| Sample imputer used | n | - |
| Feature engineering used | y | Created Health\_Score, Sleep\_Efficiency. |
| Standard deviation check | y | std() function in descriptive statistics |
| Variance check | y | var() function in descriptive statistics |
| Bias check | y | Accuracy comparison by gender groups |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Data Visualisation:

|  |  |  |
| --- | --- | --- |
| **Description** | **Y/N** | **Number and feature used** |
| Corelation Matrix | y | All numeric features from merged dataset . Typically includes: Age, Sleep Duration, Quality of Sleep, Stress Level, Physical Activity, Daily Steps, Health Score, Sleep Efficiency, Daily Usage Minutes, Engagement Rate |
| Pair plots | n | But did for the sleep dataset while performing eda |
| Data definition check | n |  |
|  |  |  |
|  |  |  |

Model selection

|  |  |
| --- | --- |
| **Description** |  |
| No of model tested | 2- Health Impact Prediction Model ,Platform Prediction Model |
| Model select criteria | GridSearchCV (accuracy) |
| Method used to select models | Traditional |
| Model accuracy | 71% |
| Confusion matrix | Actual vs predicted data of various platforms |
| R2 Value | - |
|  |  |
|  |  |

**Challenges:**

|  |  |
| --- | --- |
| Challenges | Solution |
| **Missing Values-**  Many rows have NaN values. | We removed it |
| Inconsistent Data Types | Ensures correct types for mathematical operations and visualizations. Categorical types also save memory. |
| Many duplications | We removed all duplications |
| Faced problem while merging the data as no column in two datasets were same | Somehow managed by doing it using person id.but still not overcoming. |