Mental Health Survey Dataset

by

Audwit Nafi Anam

Analysis Objective

Draw Insights on how mental health status of students differ based on
- sleep patterns,
- stress levels, and
- exercise habits

Dataset Summary
- 1053 Samples, 5 Columns
  - Age, Gender, Sleep Time, Exercise Frequency, Interaction Frequency
  - Target Label: Mental Health Status (MHS)

Dataset Preprocessing

- Deleting Rows with Null Values (26) and Duplicates (173)
- One-hot Encoding Categorical Features
  - Gender
- Label Encoding Ordinal Features
  - Mental Health Status (Poor: 0, Fair: 1, Good: 2)
  - 8 Columns after Encoding
- Outlier Removal
  - Keeping Numerical Values within 1.5*IQR
  - Keeping Categorical Values having count greater than threshold
  - 822 Samples after removal
- Challenges
  - Duplicate values with slightly different names
  - Trailing/Leading Whitespaces
- Solutions
  - Used lambda exprsssions, regular expressions, etc.

Exploratory Data Analysis (EDA)

- ~60% have fair 36% have poor and 3.5% have good mental health status (MHS)
- Sleep times range from 2 to 10 hours, averaging 6.5 hours for most
- Stress level ranges from 1-10 with mid-high average (6/10)
- Rarely exercise (average frequency 2/10)
- Moderate Social Interaction (average 4/10), most rarely interact (<4)
- Males sleep less in general (6-7), rest have wider range (6-8)
- Other Gender exercisea more (2-4)
- Higher sleep time corresponds to better MHS
- Fair MHS have low range of stress level (around 6), poor have range 4-8 and good have 3-8
- Good MHS follows higher exercise frequency
- Higher social interaction leads to better MHS

Model Development and Evaluation

- **Logistic Regression**
  - Accuracy: 68.42%
  - Macro Avg F1: 53%
- **SVM**
  - Linear Kernel:
    - Accuracy: 74%, F1: 76%
  - RBF
    - Accuracy: 62%, F1: 25%
  - Sigmoid
    - Accuracy: 63%, F1: 29%
- **Random Forest**
  - All metrics are 100%
  - The target variable MHS derived from input features; model learned the relation easily

Key Insights and Recommendations

Insights
- LR model performs decent
- SVM with linear kernel performs best – data linearly seperable
- RF model just `recognizes that condition or pattern`, and performs perfectly

Recommendations
- Gather more data
- Balanced data
- Modularize, Containerize for scalable deployment

Thank You