Cyberpsychology

(The Confluence of Technology in Mental Resilience)

23L-2637 , 23L-2516 , 23L-2560

Department of Data Science, National University of Computer and Emerging Science, Lahore.

Abstract

This project focuses on leveraging machine learning to predict mental health status, addressing a critical area of human well-being. The dataset comprises features such as stress levels, sleep hours, different types of technology usage hours, gaming activity, and physical activity, which significantly impact mental health. Three machine learning models were implemented: Logistic Regression, Random Forest, and XGBoost. Feature importance derived from these models was used for feature selection, enhancing the prediction accuracy. XGBoost achieved the highest accuracy, underscoring its effectiveness for this classification task. The findings aim to facilitate early detection and proactive management of mental health issues.

---

1. Introduction

Mental health is a fundamental aspect of human well-being, influencing individuals' thoughts, emotions, and behaviours. Accurate prediction and early detection of mental health issues are essential for promoting overall health and preventing severe outcomes. Leveraging machine learning for mental health prediction, by analysing 5000 persons' data, offers a systematic and unbiased approach to identifying at-risk individuals and facilitating timely intervention.

This study explores the factors affecting mental health using a comprehensive dataset comprising attributes such as:

- **Stress Level**: Self-reported stress scores.
- **Sleep Hours**: Average daily sleep duration.
- **Screen Time**: Hours spent using electronic devices.
- **Physical Activity Hours**: Engagement in physical activities.
- **Gaming Hours**: Time allocated to video gaming.

The research emphasizes understanding the relationships between features, identifying key predictors, and employing machine learning models to enhance prediction accuracy. Preliminary analysis involved data cleaning, visualization, and feature selection to ensure a high-quality dataset for model training. Techniques like correlation matrices and visual plots were used to highlight the most influential factors.

The aim is to identify the most effective approach by implementing methods such as Logistic Regression, Random Forest, and XGBoost for mental health prediction, contributing to proactive mental health management strategies.

2. Methodology

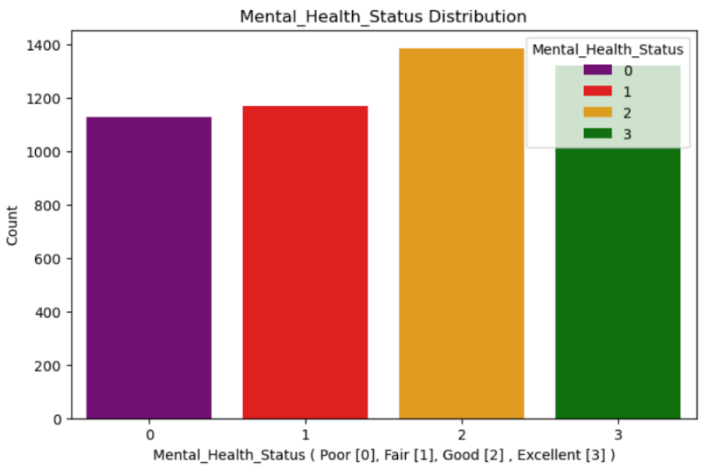
2.1 Data Preprocessing

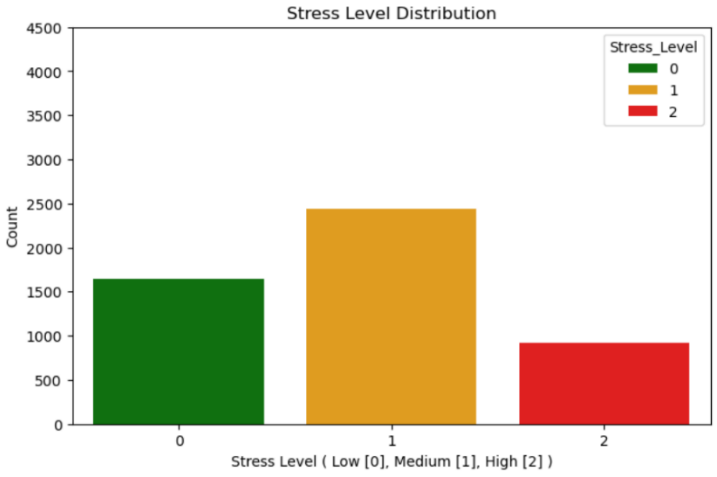
The dataset underwent the following preprocessing steps:

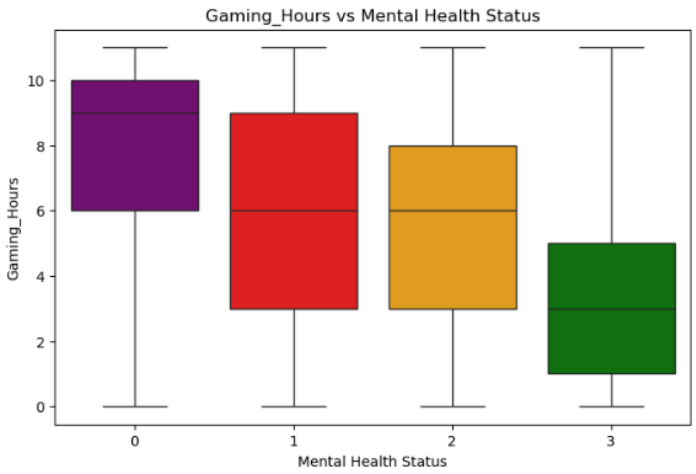
- Removal of missing or irrelevant data.
- Checking for outliers.
- Encoding of categorical data in meaningful ordinal numerical values
- Analysing each feature's important statistics and shape of data.

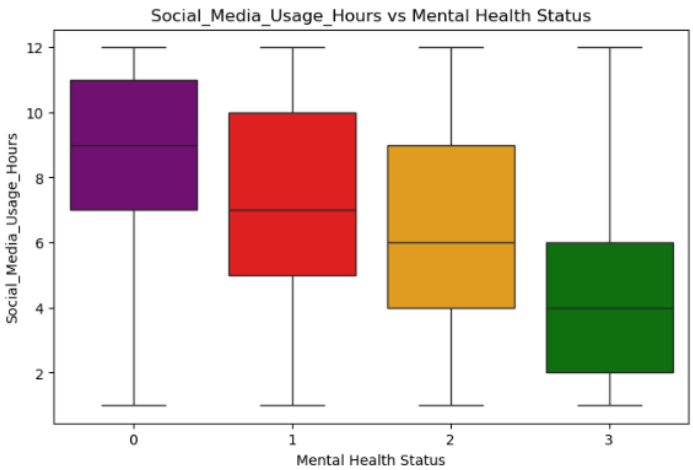
2.2 Data Visualization

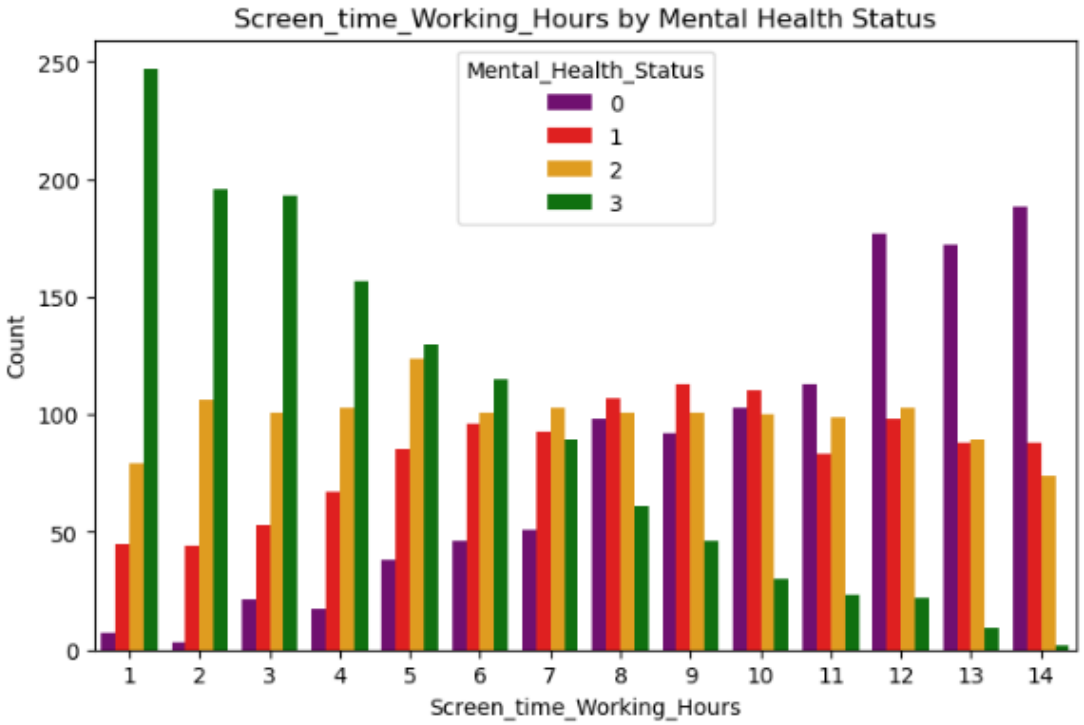
Visualizations such as heatmaps, boxplots, bar charts, pie charts etc. were employed to understand feature correlations and distributions, incorporating univariate and bivariate analysis.

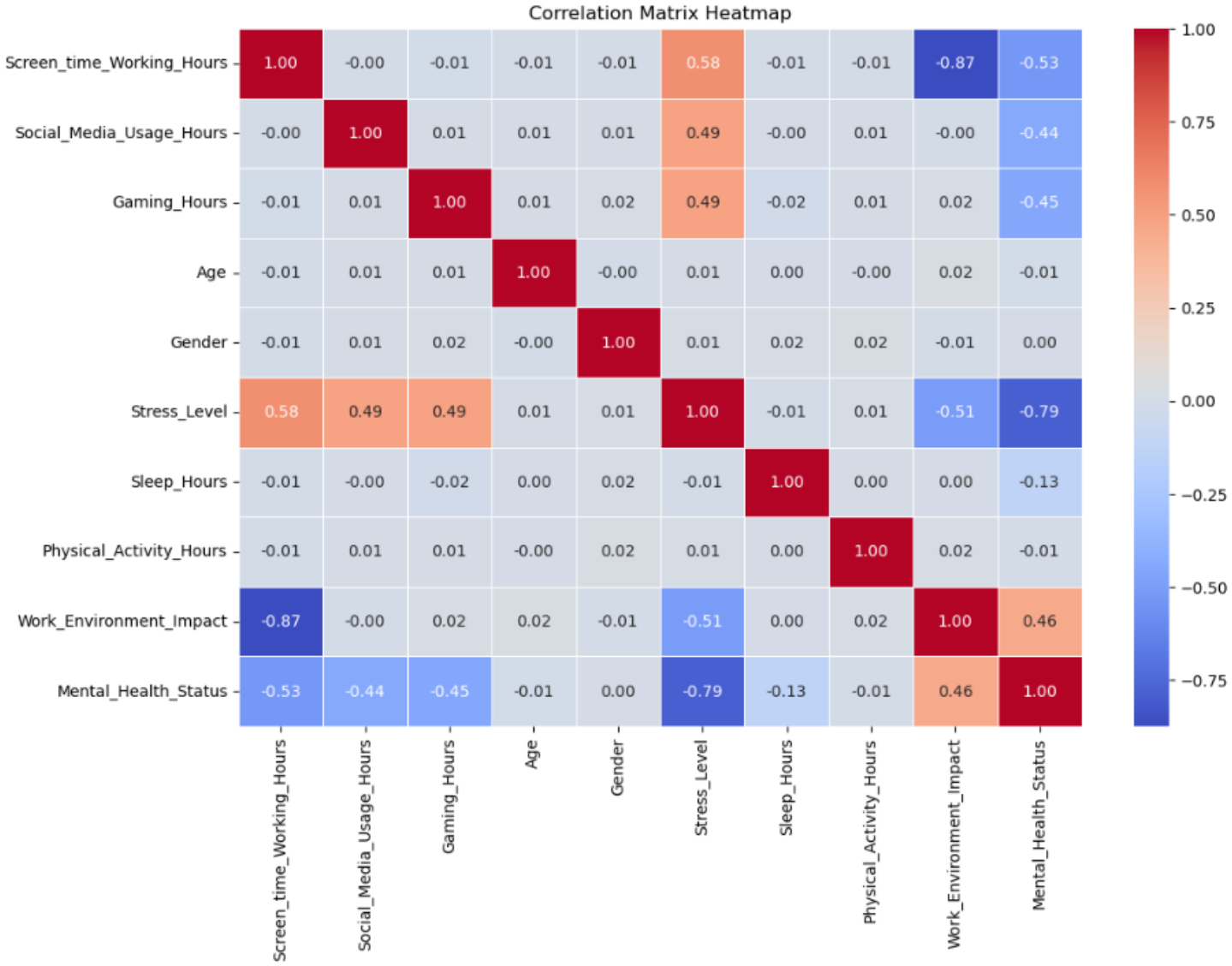









Screen_time_Working_Hours by Mental Health Status
Mental_Health_Status
0
1
2
3
Count
250
200
150
100
50
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
Screen_time_Working_Hours


Correlation Matrix Heatmap
Screen_time_Working_Hours 1.00 -0.00 -0.01 -0.01 -0.01 0.58 -0.01 -0.01 -0.87 -0.53
Social_Media_Usage_Hours -0.00 1.00 0.01 0.01 0.01 0.49 -0.00 0.01 -0.00 -0.44
Gaming_Hours -0.01 0.01 1.00 0.01 0.02 0.49 -0.02 0.01 0.02 -0.45
Age -0.01 0.01 0.01 1.00 -0.00 0.01 0.00 -0.00 0.02 -0.01
Gender -0.01 0.01 0.02 -0.00 1.00 0.01 0.02 0.02 -0.01 0.00
Stress_Level 0.58 0.49 0.49 0.01 0.01 1.00 -0.01 0.01 -0.51 -0.79
Sleep_Hours -0.01 -0.00 -0.02 0.00 0.02 -0.01 1.00 0.00 0.00 -0.13
Physical_Activity_Hours -0.01 0.01 0.01 -0.00 0.02 0.01 0.00 1.00 0.02 -0.01
Work_Environment_Impact -0.87 -0.00 0.02 0.02 -0.01 -0.51 0.00 0.02 1.00 0.46
Mental_Health_Status -0.53 -0.44 -0.45 -0.01 0.00 -0.79 -0.13 -0.01 0.46 1.00
Screen_time_Working_Hours
Social_Media_Usage_Hours
Gaming_Hours
Age
Gender
Stress_Level
Sleep_Hours
Physical_Activity_Hours
Work_Environment_Impact
Mental_Health_Status
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

2.3 Feature Selection

Instead of traditional dimensionality reduction techniques, feature importance metrics from models were utilized. Unimportant features like gender and work environment impact were removed, resulting in improved model accuracy.

---

3. Experiments

Three models were implemented:

Logistic Regression

Initial accuracy: 74.13% with the following features:

```
Feature Importances:
                      Feature  Importance
5               Stress_Level    0.246569
6                Sleep_Hours    0.232500
2               Gaming_Hours    0.126878
1   Social_Media_Usage_Hours    0.121342
0  Screen_time_Working_Hours    0.107158
3                        Age    0.070143
7    Physical_Activity_Hours    0.046501
8    Work_Environment_Impact    0.034495
4                     Gender    0.014414
```

After feature selection: 74.13% (No change)

Random Forest

Initial accuracy: 91.7% with following features:

```
Feature Importance in Logistic Regression:
                      Feature  Importance
0  Screen_time_Working_Hours    3.349372
2               Gaming_Hours    2.831285
1   Social_Media_Usage_Hours    2.684488
6                Sleep_Hours    1.105741
5               Stress_Level    0.478121
3                        Age    0.042718
7    Physical_Activity_Hours    0.033497
4                     Gender    0.020459
8    Work_Environment_Impact    0.001793
```

After feature selection: 92.2%

XGBoost (Final Model)

Initial accuracy: 94.8%

```
Feature Importances:
                      Feature  Importance
5               Stress_Level    0.853851
6                Sleep_Hours    0.068384
1   Social_Media_Usage_Hours    0.023998
2               Gaming_Hours    0.020596
0  Screen_time_Working_Hours    0.019263
3                        Age    0.005133
7    Physical_Activity_Hours    0.004667
4                     Gender    0.004107
8    Work_Environment_Impact    0.000000
```

After feature selection: **95.4%**

---

4. Results & Discussion

XGBoost consistently outperformed Logistic Regression and Random Forest in precision, recall, and overall accuracy. Feature selection improved model performance, validating the utility of model-driven feature importance.

```
Accuracy: 0.954
Confusion Matrix:
 [[230   8   0   0]
 [  2 206   8   0]
 [  0  14 276   8]
 [  0   0   6 242]]
Classification Report:
              precision    recall  f1-score   support

           0       0.99      0.97      0.98       238
           1       0.90      0.95      0.93       216
           2       0.95      0.93      0.94       298
           3       0.97      0.98      0.97       248

    accuracy                           0.95      1000
   macro avg       0.95      0.96      0.95      1000
weighted avg       0.95      0.95      0.95      1000

Feature Importances:
                      Feature  Importance
4               Stress_Level    0.854368
5                Sleep_Hours    0.069861
1   Social_Media_Usage_Hours    0.024917
2               Gaming_Hours    0.020786
0  Screen_time_Working_Hours    0.020075
3                        Age    0.005284
6    Physical_Activity_Hours    0.004709
```

R^2 (coefficient of determination): 0.9610

---

5. Conclusion

The findings demonstrate the potential of machine learning to address mental health challenges. XGBoost, with its superior accuracy, is recommended for such predictive tasks. Future studies could explore integrating more behavioral data and applying advanced ensemble techniques to enhance predictions further.