Kyoungmi Kwon

Urban Mouse vs. Rural Mouse

Testing the Modern Relevance of Aesop's Fable Using Mental Health Data from Real World

1. Abstract

This project explores whether the core message of Aesop's fable "The Town Mouse and the Country Mouse" still holds in a modern context, particularly regarding mental health disparities between urban and rural populations. Using a Kaggle dataset which is with cities in India, I tested hypotheses about how living environments influence stress, life style, and mental health conditions. Results show that while some traditional assumptions are challenged, key themes of the fable - such as the trade-off between opportunity and peace of mind - are still relevant.

This project was developed as part of a bootcamp datathon challenge between May 30 and June 3, 2025. The goal was to apply and demonstrate skills learned throughout the program by conducting an end-to-end data analysis project using a Kaggle dataset. The project was completed as a team project with shared responsibilities in data visualization, and others.

2. Background and Motivation

In Aesop's fable, the urban mouse enjoys luxury but lives in fear and discomfort, while the rural mouse lives a simpler but safer life. This moral story inspired the hypothesis that modern urban life may be associated with higher

stress and worse mental health outcomes, despite better infrastructure and opportunity. This project aims to test this claim empirically using real-world data.

The idea for this project emerged while observing the datathon dataset. Our team included senior participants in their late 60s, so I looked for a topic that would be easy to grasp without requiring deep domain knowledge. Upon discovering that the dataset contained mental health indicators across Indian cities, Aesop's fable of the Town Mouse and The Country Mouse immediately came to mind.

India's unique social fabric - where urbanization, wealth gaps, and traditional rural life coexist- seemed like an ideal setting to test the moral of the fable in a modern context. The nation's rapid urban growth, paired with persistent regional and economic disparities, provided a rich foundation for examining how living environments influence mental wellbeing.

3. Data Source

- Dataset : Exploring Mental Health Data
  https://www.kaggle.com/competitions/playground-series-s4e11/overview
- Number of observations : 51,880 entries
  - Urban(Top 7 cities by GRP) : 26,808
  - Rural (Bottom 5 cities by GRP) : 25,072
- Variables include :
  - Lifestyle Indicators :
    - Sleep Duration
    - Dietary Habits
    - Work/Study Hours
  - Stress Indicators :
    - Academic Pressure
    - Work Pressure
    - Financial Stress

- Mental Health Indicators :
  - Depression
  - Have you ever had suicidal thoughts
  - (Family History of Mental Illness as a background factor for mental health)
- Demographic/Contextual :
  - Age (grouped into age brackets)
  - Degree
  - Working Profession
  - Profession
  - City (used for regional classification)
- Other : CGPA (academic score)
- Urban/Rural Classification Method:

  Cities were ranked by per capita Gross Regional Product (GRP), sourced from Indian public statistics. The top 7 cities were labeled as. “Urban”, while the bottom 5 cities were grouped as “Rural”. This approach ensured regional and socioeconomic diversity while keeping the class sizes balanced.

4. Research Questions and Hypotheses

RQ1: Do urban and rural populations in India differ significantly in lifestyle factors?

RQ2: Is stress level higher in urban areas?

RQ3: Are mental health risks such as depression or suicidal thoughts more prevalent in urban environments?

Hypothesis 1: Urban residents experience poorer sleep, diet, activity balance.

Hypothesis 2: Urban residents have higher levels of psychological stress.

Hypothesis 3: Urban populations are at higher risk for mental health issues.

5. Methodology

- Data Preprocessing :
  - All comparisons were made by analyzing 100% ratios for both population to account for sample size differences.
  - Converted numerical variables into categorical bins using relevant domain standards :
    - Sleep duration :
      Categorized based on sleep medicine standard optimal = 7-8 hours, insufficient < 7, excessive > 8
    - Work/Study Hours :
      Grouped into : insufficient < 4h, optimal 4-8h, slightly excessive 8-10h, excessive >12h
    - Age : Grouped into age bracket
  - Merged 'Profession' and 'Working Status' columns; removed conflicting entries
  - Recategorized profession types into 10 major categories
  - One-Hot encoding applied to categorical variables where needed
  - Urban/Rural labeling added as a new categorical variable
  - Imbalanced sample size considered when comparing urban and rural groups
- Statistical analysis :
  - Chi-square tests:
    Assessed whether distributions of demographic and lifestyle features differed significantly between urban and rural populations
  - Spearman Correlation:
    Measured association between mental health outcomes (d.g., depression, suicidal thoughts) and stress/lifestyle indicators.

- Machine Learning Models:
  Predict whether a respondent is likely to experience depression (binary classification)
  - Logistic Regression
  - Random Forest
  - XGBoost, LightGBM
- Performance metric : Accuracy
  All models achieved over 93% accuracy on the test set
- Feature Importance:
  Used to identify which factors more significantly affect mental health, Top predictors included financial stress, profession type, and demographic features.

5 - 1. Composite Scoring Strategy

To enable intuitive comparisons between urban and rural populations, three thematic composite scores were developed and normalized to a 0-1 range:

- Lifestyle Score

  Calculated based on three categorical variables : sleep duration, dietary habits, and work/study hours.

  Each feature was scored based on domain-recommended standards and aggregated into a total of 6 points, which was normalized to a 1-point scale.

  Higher values indicate healthier lifestyle patterns.

- Stress Score

  Derived from three indicators: work/study pressure, financial stress, and work/study satisfaction.

  The satisfaction variable was inverted and combined with the other two to produce a total score between -5 and +10, which was then normalized to a 1-point scale.

  Higher values indicate greater stress.

- Mental Health Score

Based on binary indicators of depression and suicidal thoughts, each contributing 1 point if present (total max = 2)

The score was normalized to 1 point.

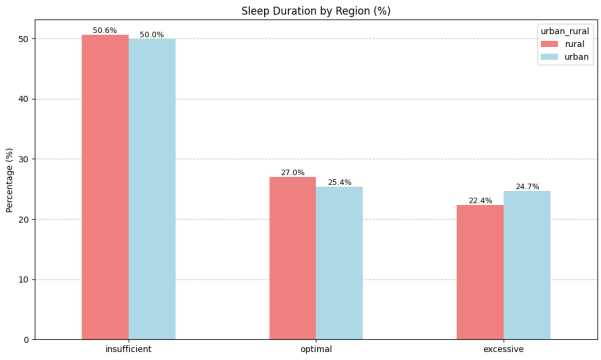
Higher values indicate poorer mental health conditions.

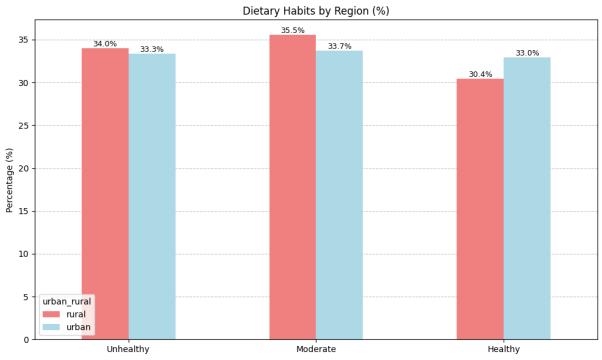
Score Type	Rural	Urban	Reference
Life_style_score	0.4850	0.4866	Higher is better
Stress_score	0.5340	0.5413	Lower is better
Mental_health_score	0.3350	0.3450	Lower is better

These scores helped consolidate multiple indicators into interpretable indices and allowed for easier comparison across demographic groups.

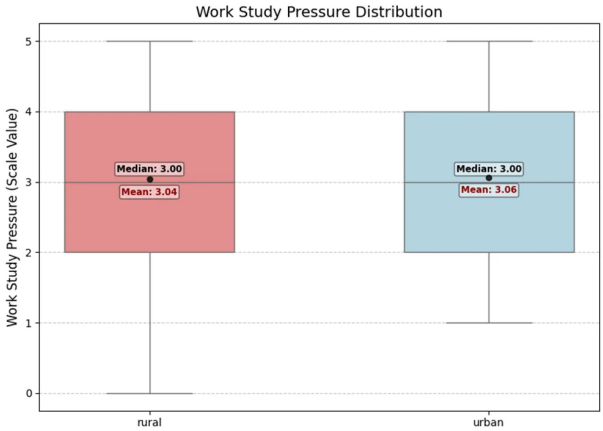
6. Key Results

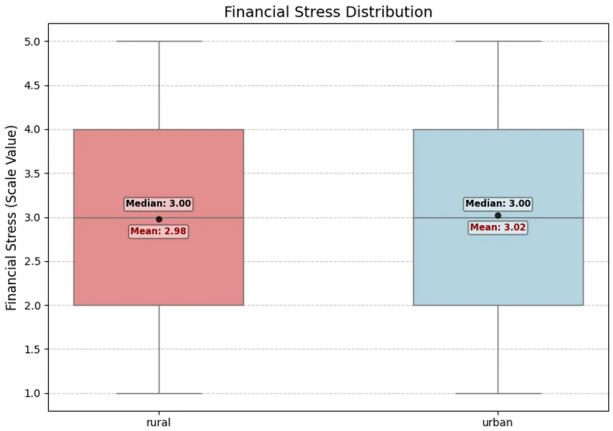
- Sleep and diet patterns showed minimal statistically significant differences.



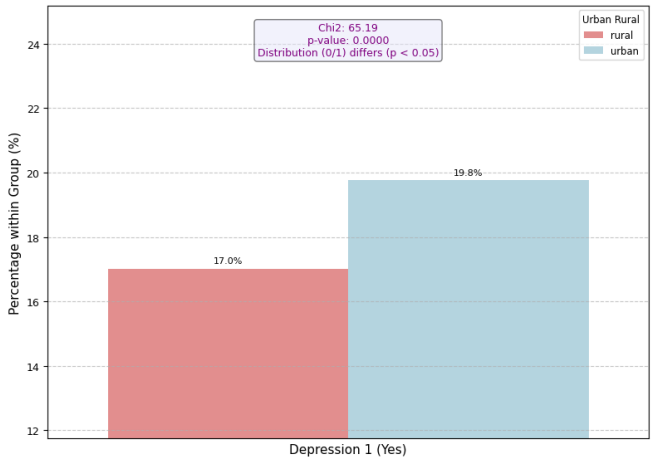


- Stress levels were higher among urban respondents, related to financial and academic/work pressure.

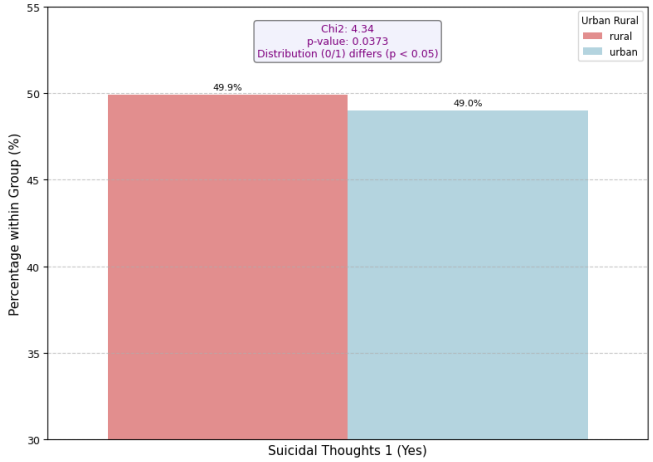




- Depression were more prevalent among urban respondents.



- Suicidal thoughts were more prevalent among rural respondents.



- Urban residents have similar scores to rural residents, but have slightly higher stress scores and slightly worse mental health.

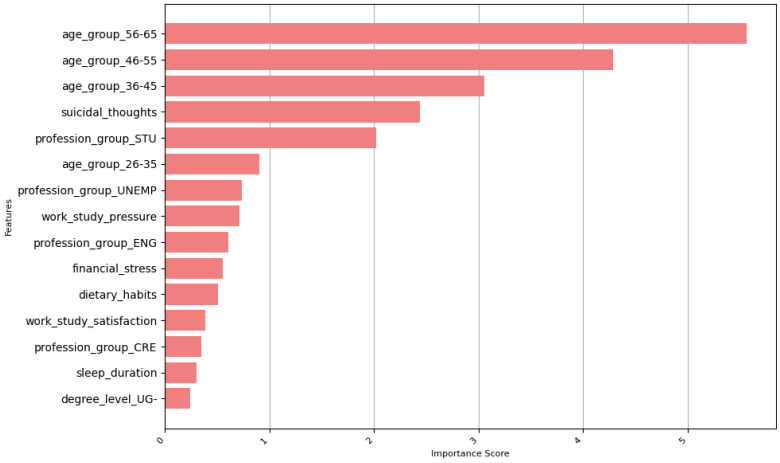
Score Type	Rural	Urban	Reference
Life_style_score	0.4850	0.4866	Higher is better
Stress_score	0.5340	0.5413	Lower is better
Mental_health_score	0.3350	0.3450	Lower is better

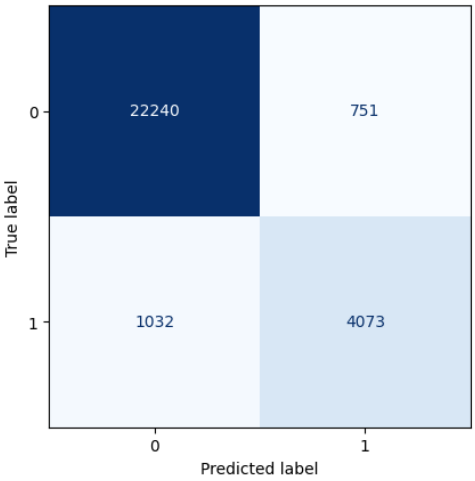
"Table 1. Composite Score Comparison between Urban and Rural Groups"

- All models achieved over 93% accuracy in predicting mental health risks.

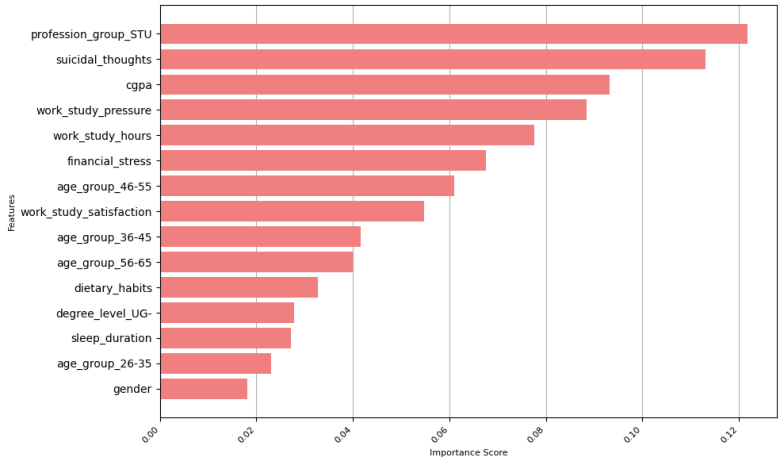
- Stress indicators and demographic characteristics(e.g., age, profession) as the most influential factors impacting mental health.

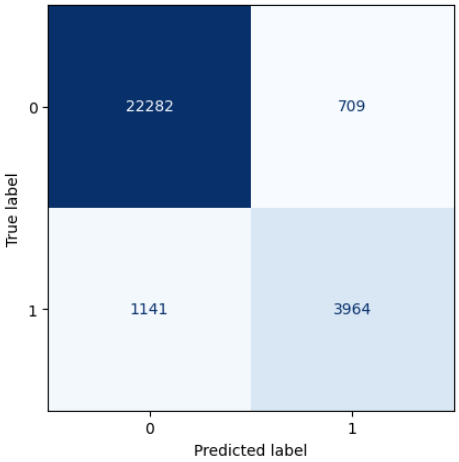
<Logistic Regression>



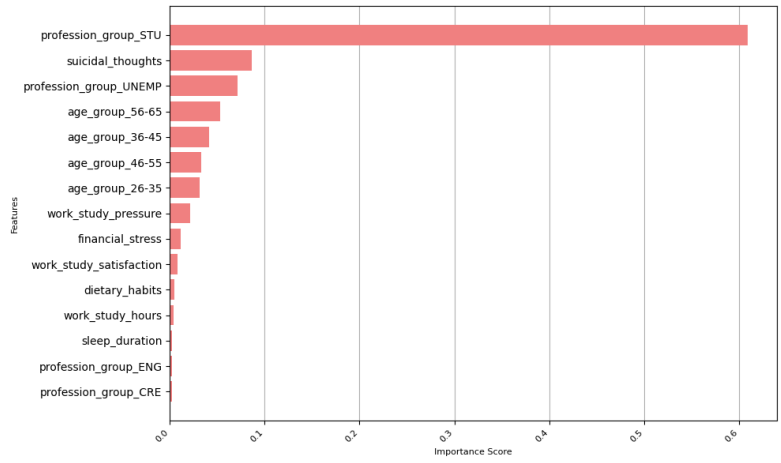


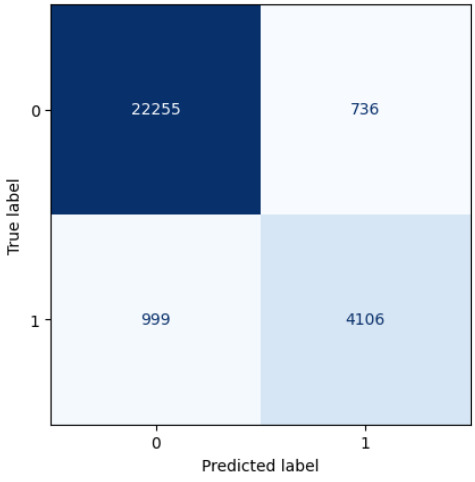
<Random Forest>





<XGBoost>





7. Interpretation

The results partially support the fable's message : while urban residents may have better access to resources, they also experience greater psychological strain. However, some lifestyle habits such as diet were actually better among urban respondents, possibly due to education or income.

This calls for a more nuanced understanding of modern mental health dynamics in the context of rapid urbanization

8. Limitation and Future Work

Dataset Specificity : The dataset is self-reported, which can introduce bias. And the findings are based on a specific dataset from India, limiting generalization to other countries or regions.

Variable Granularity : The dataset's variable granularity and measurement methods might have limited the ability to find strong direct correlations between certain lifestyle/stress factors and mental health indicators.

Data sampling & Matching Challenges : one of the key challenges was the inability to extract urban and rural samples with perfectly matched demographic characteristics(e.g., age, education). As a result, comparisons were made based on overall population distributions rather than controlled subgroup matching, which may have introduced confounding effects.

Future Research : We recommend future studies to involve more in-depth surveys specifically designed to explore causal relationships between lifestyle, socio-economic factors, and mental health, possibly with a focus on specific regional industrial/economic contexts.

9. Technologies

- Python
- Pandas, Numpy
- Seaborn, Matplotlib
- Scikit-learn

- XGBoost
- Jupyter Notebook
- GitHub : https://github.com/KyoungmiKwon/fable-validation-mental-health

10. Conclusion

This project demonstrates how a classical narrative can be tested using empirical methods. The fable's lesson still resonates: modern comfort does not guarantee mental peace. For policymakers and researchers, these findings stress the importance of understanding the psychological cost of urbanization.