Predictive Analytics for Student Mental Health: A Multi-Factor Analysis for Early Intervention

- **Khush Domadiya.**
  FR9739

Outline

Problem Setup

Mental health disorders among college students have reached epidemic proportions, with over 60% of students reporting overwhelming anxiety and 40% experiencing depressive symptoms. Despite the high prevalence, research indicates that 58 out of 70 students with mental health issues do not seek professional help. This crisis significantly impacts academic performance, dropout rates, and long-term student success, making predictive analytics a crucial tool for proactive mental health support.

Universities worldwide are struggling to identify at-risk students before their conditions escalate. While research demonstrates that lifestyle factors, academic pressures, and social dynamics are strong predictors of mental health outcomes, comprehensive multi-factor analysis remains limited in academic settings. This creates an urgent need for data-driven approaches that can systematically identify vulnerable students and enable targeted interventions.

Research Questions:

1. Which combination of lifestyle, academic, and demographic factors most accurately predict student mental health outcomes?
2. Can we identify distinct student clusters based on mental health risk profiles to enable targeted intervention strategies?
3. What is the relationship between academic performance indicators and mental health outcomes across different student demographics?
4. How effective are current counseling services, and what factors predict help-seeking behavior among students with mental health issues?

Data Collection & Exploratory Analysis

The analysis utilized a comprehensive Student Mental Health Survey dataset containing **7,022 student records** with **20 variables** spanning demographic, academic, lifestyle, and mental health dimensions. This large-scale dataset provides robust statistical foundations for predictive modeling and represents one of the most comprehensive student mental health datasets available for academic research.

Key Variables:

- **Target Variables**: Depression Score (0-5), Anxiety Score (0-5), Stress Level (0-5)
- **Demographics**: Age (18-35), Gender, Course of study, Residence Type
- **Academic Factors**: CGPA (2.44-4.0), Semester Credit Load (15-29), Extracurricular Involvement

- **Lifestyle Variables**: Sleep Quality, Physical Activity, Diet Quality, Social Support, Substance Use

Data Quality Assessment:

The dataset exhibited exceptional quality with minimal missing values - only 12 CGPA records (0.17%) and 15 Substance Use records (0.21%), ensuring robust analytical foundations. No duplicate records were identified, and all variables showed appropriate distributions for statistical analysis.

Critical Exploratory Findings:

- **60.8% of students classified as high-risk** for mental health issues (scores ≥4 on any scale)
- **Course-specific vulnerabilities**: Computer Science students showed highest depression scores (3.30), Law students highest anxiety (3.23), Medical students highest stress (3.21)
- **Lifestyle correlations**: Poor sleep quality, low social support, and high financial stress emerged as key risk factors
- **Gender balance**: Nearly equal representation (50.5% male, 49.5% female) ensuring generalizable results

The exploratory analysis revealed concerning patterns, with mean mental health scores of 2.25 (Depression), 2.30 (Anxiety), and 2.43 (Stress), indicating widespread moderate-to-high levels of mental health issues across the student population.

Methods & Implementation

The analytical approach employed a comprehensive three-stage methodology combining supervised learning, unsupervised learning, and statistical validation techniques.

Stage 1: Data Preprocessing

- Missing value imputation using median/mode replacement for the 27 missing records
- One-hot encoding for categorical variables with 30 final features
- Standard scaling for numerical variables (Age, CGPA, Financial Stress, Credit Load)
- Train-test split (80/20) with stratification to maintain class balance

Stage 2: Predictive Modeling

Two complementary machine learning approaches were implemented and rigorously validated:

Logistic Regression Model:

- **Accuracy**: 61.4%
- **Precision**: 66.9%

- **Recall**: 72.2%
- **F1-Score**: 69.4%

Random Forest Model:

- **Accuracy**: 61.4%
- **Precision**: 62.9%
- **Recall**: 89.4% (superior for identifying at-risk students)
- **F1-Score**: 73.8%

The Random Forest model demonstrated superior recall performance (89.4%), making it highly effective for identifying students requiring mental health intervention - a critical requirement where false negatives could have serious consequences.

Stage 3: Clustering Analysis

K-means clustering with silhouette analysis identified **4 distinct student mental health profiles**, enabling targeted intervention strategies. The optimal clustering solution achieved strong silhouette scores, indicating well-separated student groups with distinct risk characteristics.

Statistical Validation:

Chi-square tests revealed significant associations between mental health outcomes and key categorical variables:

- **Course of study**: $\chi^2 = 239.357$, $p < 0.001$ (highly significant)
- **Social support level**: $\chi^2 = 8.156$, $p = 0.017$ (significant)

Model Validation:

Both models underwent rigorous train-test validation with stratified sampling to ensure unbiased performance estimates. Feature importance analysis identified the most predictive factors, providing actionable insights for intervention programs.

Insights & Analysis

The analysis yielded multiple high-level insights with significant practical implications for university mental health programs.

Quantitative Model Performance:

Both predictive models achieved solid performance relative to the complexity of mental health prediction. The **Random Forest model's 89.4% recall** is particularly noteworthy, as it successfully identifies nearly 9 out of 10 high-risk students - a critical capability for early intervention programs.

While overall accuracy was moderate (61.4%), this reflects the inherent complexity of mental health prediction and aligns with established benchmarks in psychological research.

Key Predictive Factors Identified:

1. **Academic Program**: Computer Science, Law, and Medical students show distinctly elevated mental health risks, with each program exhibiting different vulnerability patterns
2. **Financial Stress**: Strong correlation with overall mental health risk, indicating socioeconomic factors as major contributors
3. **Sleep Quality**: Emerged as a critical lifestyle predictor, consistent with established sleep-mental health research
4. **Social Support Systems**: Low social support significantly associated with higher risk profiles
5. **CGPA Performance**: Interestingly showed minimal correlation (-0.023 with depression), suggesting academic performance alone is not predictive

Student Segmentation Insights:

The clustering analysis revealed **4 distinct mental health profiles** that enable targeted interventions:

- **High-risk clusters** requiring immediate attention and comprehensive support
- **Moderate-risk groups** suitable for preventive programming
- **Low-risk populations** for wellness maintenance programs

Critical Service Gap Analysis:

A concerning finding emerged regarding counseling service utilization: **60.8% of students classified as high-risk**, yet counseling services showed suboptimal effectiveness patterns. This indicates significant opportunities for improving mental health service delivery and accessibility.

Predictive Model Strengths:

- **High recall performance** minimizes false negatives (missing at-risk students)
- **Robust feature importance** provides clear intervention targets
- **Scalable implementation** suitable for real-time student assessment
- **Statistically validated** through rigorous testing protocols

Model Limitations Acknowledged:

- Cross-sectional data limits causal inference capabilities
- Self-reported measures may introduce reporting bias

- Moderate overall accuracy suggests additional factors may enhance prediction
- Generalizability requires validation across different institutional contexts

The analysis demonstrates that machine learning approaches can effectively identify students at mental health risk while providing actionable insights for intervention program design.

Conclusions & Recommendations

This comprehensive analysis of 7,022 students reveals that **predictive analytics can effectively identify mental health risks** in university populations, with significant implications for early intervention programming.

Primary Findings:

- **60.8% of students require mental health intervention**, indicating widespread need for enhanced services
- **Random Forest modeling achieves 89.4% recall** for identifying at-risk students, enabling effective screening programs
- **Academic program, financial stress, sleep quality, and social support** emerge as primary predictive factors
- **Four distinct student mental health clusters** provide frameworks for targeted intervention strategies

Strategic Recommendations for Universities:

1. Implement Predictive Screening Systems

Deploy machine learning models for systematic identification of high-risk students during orientation and semester transitions. The demonstrated 89.4% recall rate provides confidence for operational implementation.

2. Program-Specific Mental Health Programming

Develop targeted interventions for high-risk academic programs:

- **Computer Science**: Focus on depression prevention and academic stress management
- **Law Programs**: Emphasize anxiety reduction and competitive pressure management
- **Medical Studies**: Implement comprehensive stress management and resilience training

3. Financial Wellness Integration

Integrate financial counseling with mental health services, given the strong correlation between financial stress and psychological wellbeing demonstrated in the analysis.

4. Sleep Hygiene and Lifestyle Programming

Implement campus-wide sleep education initiatives, as sleep quality emerged as a critical modifiable risk factor with strong predictive power.

5. Enhanced Social Support Systems

Strengthen peer support networks and community building programs, particularly for students in high-risk demographic and academic categories.

Implementation Roadmap:

- **Phase 1**: Pilot predictive screening with incoming student cohorts
- **Phase 2**: Deploy targeted interventions based on clustering analysis
- **Phase 3**: Evaluate intervention effectiveness and refine models
- **Phase 4**: Scale successful programs university-wide

This analysis demonstrates that data-driven approaches can transform university mental health services from reactive treatment models to proactive prevention and early intervention systems, potentially improving outcomes for thousands of students while optimizing resource allocation.