Case Study: Predicting Mental Health Patterns in College Students

Prepared by: Aayush Tiwari
Project Context: Independent Submission for Harvard University
Date: July 14th

Introduction

Mental health concerns among college students have become a growing issue in recent years, with rising levels of stress, anxiety, and depression. These conditions, if left unidentified, can significantly impact students' academic performance and personal well-being. The aim of this project was to design and develop an end-to-end data-driven solution that could accurately predict students at mental health risk using their behavioral, psychological, and academic profiles. This study was undertaken independently and aligned with standards expected for research submitted to a university like Harvard, although I am not a Harvard graduate myself.

Project Objective

The primary objective was to create a system that can identify at-risk students early and assist educational institutions in offering proactive mental health support. The approach covered the entire analytics lifecycle, beginning with raw data management in Excel, continuing through data science processes using Python, and concluding with visual communication through Streamlit applications, Power BI dashboards, and business-level documentation. The final deliverables were prepared for both technical and non-technical audiences.

Data Preparation

The dataset was initially handled in Microsoft Excel, where raw data included student ID, name, gender, academic year, sleep hours, perceived stress score, study intensity, therapy history, and depression score. Several data preprocessing steps were conducted. Feature engineering techniques were applied to convert sleep hours into categorized sleep quality bands, and stress scores were binned into low, moderate, and high categories. Additional categorical variables were derived for analytical clarity. Pivot tables were created to explore patterns such as the percentage of students at risk by academic year, the relationship between sleep quality and stress, and the average depression score in relation to therapy history.

Model Development in Python

The data was then imported into Jupyter Notebook, where a machine learning model was developed using the Random Forest Classifier. The model pipeline included encoding categorical variables, scaling numerical features, and splitting the data into training and testing sets. GridSearchCV was used to fine-tune hyperparameters and improve model performance. Surprisingly, the final model achieved 100% accuracy, precision, recall, and F1-score on the test set. While such results are rare and impressive, they can also indicate the potential for overfitting or the presence of highly separable patterns in the dataset. This observation was

noted with caution, and the need for testing on more generalized real-world data was acknowledged.

Deployment and Visualization

To make the solution accessible to users, a Streamlit application was built, allowing interactive input of student attributes and providing immediate predictions on whether a student is likely at risk. The user interface was designed to be intuitive for non-technical users such as academic counselors or student support officers. In parallel, Power BI was used to develop an interactive dashboard that visualized key metrics such as stress distribution, risk percentage across academic years, and depression trends in students who have or haven't attended therapy. DAX measures were written to compute actionable KPIs such as average depression score and stress levels by demographic segments.

Business Relevance

From a strategic perspective, this solution equips educational institutions with a framework to proactively identify students in need of support. It has the potential to improve retention rates, reduce academic failures, and enhance student life outcomes by enabling early interventions. Counseling teams can use the dashboard to plan outreach efforts, and university leadership can leverage model outputs for long-term planning. By translating technical predictions into visually accessible insights, the project bridges the gap between data science and decision-making.

Limitations and Recommendations

While the 100% model performance suggests strong correlations in the dataset, it also raises the possibility that the dataset may be limited or too clean. Further testing on larger, noisier datasets is recommended to verify generalization capability. Additionally, ethical considerations around student privacy and mental health labeling must be considered when deploying such models in production.

Conclusion

This end-to-end project demonstrates how data science techniques can be effectively applied to a socially impactful domain like student mental health. Through systematic data handling, intelligent modeling, and thoughtful presentation, a working prototype was created that not only predicts risk accurately but is also deployable and understandable by institutional stakeholders. Although conducted independently, the quality and intent of this project align with standards of rigorous academic research, such as those expected by Harvard University. The solution holds promise for real-world implementation in educational settings, and further development could lead to a powerful tool for institutional well-being management.