Predicting Depression in University Students: A Machine Learning Approach

Executive Summary

This project aimed to address the business question: *Can we predict student depression based on survey data to allow earlier intervention and targeted support in a university setting?* Leveraging machine learning models and psychological insights, I developed and evaluated several predictive algorithms to identify students at higher risk of depression.

Using a dataset composed of over 29,000 student responses with variables including demographics, academic pressure, sleep patterns, suicidal thoughts, and family history of mental illness, I preprocessed, analyzed, and modeled the data to extract key predictors and build reliable classifiers. The outcome not only achieves high predictive performance (AUC ~0.928) but also yields insights into which features most impact depression risk, thus aiding university stakeholders in implementing data-driven mental health strategies.

Approach & Methodology

The workflow consisted of the following steps:

1. **Exploratory Data Analysis (EDA)**: Relationships between depression and key variables like gender, sleep, and suicidal thoughts were visualized.
2. **Preprocessing & Encoding**: Categorical variables were encoded using one-hot and ordinal encoding. Missing values were addressed, and numerical features were scaled.
3. **Modeling**: Six machine learning models were trained and evaluated:
    - Logistic Regression
    - K-Nearest Neighbors (KNN)
    - Support Vector Machine (SVM)
    - XGBoost
    - Decision Tree
    - Random Forest

    Hyperparameter tuning was performed on each model for optimization. Performance was evaluated using metrics including accuracy, precision, recall, F1 score, and ROC-AUC.
4. **Feature Importance**: Post-modeling, feature importances were extracted to determine which factors had the greatest impact on predicting depression.
5. **Hypothesis Testing**: Statistical tests were conducted to validate psychological assumptions about depression predictors.

Business Problem Solved

The central business goal was to assist educational institutions in **proactively identifying students at risk of depression**, allowing for earlier and more personalized interventions. My final models—particularly XGBoost and SVM—demonstrated excellent predictive power (ROC AUC ~0.928). These models can be used to:

- Inform university counseling services about students who may require outreach
- Tailor support services according to specific student risk profiles
- Incorporate automated risk detection into student portals or support systems

Feature analysis highlighted **Suicidal Thoughts**, **Academic Pressure**, and **Financial Pressure** as top predictors of depression, allowing for more focused interventions in areas with the highest impact.

Hypothesis Testing Results

Three hypotheses were tested using chi-square tests of independence:

1. **Gender as a Predictor**:
    - Statistically non-significant (p > 0.69)
    - Female students were not more likely to report depression
2. **Suicidal Thoughts as a Predictor**:
    - Strong and significant association (p = 0.00)
    - Students who experienced suicidal thoughts were far more likely to report depression
3. **Family History of Mental Illness**:
    - Also significant association (p = 0.00)
    - Students with such history had higher depression rates

These findings largely support the validity of using psychological theory to inform machine learning predictors. No differences in gender may be due to unknown variables such as family pressure affecting males more or the anonymous nature of the study reducing the under reporting in males.

Model Comparison

- **XGBoost** emerged as the **best overall model**, with the highest accuracy and precision, and equal F1-score compared to others.
- **SVM** delivered the **highest recall (0.883)**, making it ideal for minimizing false negatives.
- **Random Forest** was a close runner-up across all metrics.

All three top models (XGBoost, SVM, Random Forest) showed very similar performance, with only marginal differences in metrics.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.853	0.856	0.882	0.869
K-Nearest Neighbors	0.832	0.827	0.879	0.852
Support Vector Machine	0.853	0.856	**0.883**	0.869
XGBoost	**0.854**	**0.858**	0.881	0.869
Decision Tree	0.829	0.866	0.815	0.840
Random Forest	0.853	0.857	0.880	0.868

Conclusions & Recommendations

This project confirms that **machine learning can accurately predict depression risk in students using survey data**. XGBoost, my best-performing model, should be integrated into the university's digital infrastructure to flag at-risk students.

Key recommendations:

- **Deploy the model via internal systems** for regular screening
- **Target interventions** toward students with high academic pressure, high financial pressure, and especially suicidal ideation
- **Conduct annual surveys** using the same format to update and retrain models

Optional Extensions

- **Clustering** showed potential subgroups based on sleep patterns, which could guide differentiated care strategies.