**Final Report – Understanding Depression Among University Students**

*DSA 210 – Muhsin Efe Tarım – 32230*

**Introduction**

Depression is a growing concern within student populations, affecting academic performance, daily functioning, and overall well-being. Despite its prevalence, data-driven insights into its predictors are limited. In this project, we sought to understand how academic, behavioral, and socio-economic factors contribute to depression among university students. We used two complementary datasets to ensure findings were consistent across samples and employed a full data science pipeline, from data cleaning to machine learning modeling.

Chronologically, this project progressed through three distinct phases, each documented in separate reports. The initial proposal, documented in the README.md, outlined the selected datasets and the planned approach to data collection and analysis. The second report (Report.docx) focused on data collection, exploratory data analysis, and hypothesis testing, forming the statistical foundation of the project. Finally, the third report (ApplyingML\_Report.docx) concentrated on applying machine learning techniques to model depression outcomes. Each of these documents examines the corresponding stages in far greater detail than the final report, which provides a synthesized overview of the entire process.

**Project Goals**

1. Identify key depression risk factors such as sleep duration, academic achievement, and financial support.
2. Validate findings across two independent student survey datasets.
3. Apply both statistical inference and machine learning to analyze depression scores and status.
4. Provide interpretable results that may inform student wellness programs and mental health policy.

**Datasets**

**1.**

**uni\_depression.csv**

* Captures PHQ-9 scores and key academic indicators (e.g., CGPA range, scholarship status).
* Focused on understanding academic performance in relation to depression.

**2.**

**second\_Student\_Depression\_Dataset.csv**

* Offers richer behavioral and psychological data, including numeric CGPA, sleep duration, and stress levels.
* Enables broader analysis of lifestyle and socio-emotional correlates.

**Exploratory Data Analysis**

We used histograms, scatterplots, violin plots, KDEs, boxplots, and heatmaps to uncover trends.

* **Depression Scores**: Most students fall into mild-to-moderate depression categories.
* **Sleep Duration**: Students with <5 hours of sleep reported the highest depression scores.
* **CGPA**: Weak negative correlation with depression. Higher academic performance does not strongly predict lower depression.
* **Scholarship Status**: Surprisingly, recipients exhibited slightly higher depression, potentially due to stress from performance pressure.
* **PHQ-9 Item Analysis**: Core emotional symptoms (e.g., hopelessness) were most predictive of total score.

**Hypothesis Testing**

|  |  |  |
| --- | --- | --- |
| **Hypothesis** | **Test** | **Result** |
| Sleep Duration vs. Depression | Spearman &  ANOVA | Significant negative correlation (ρ = –0.087),  F = 93.01, p ≪ 0.001 |
| Family History vs. Depression | Chi-square | Strong association, χ² = 79.43, p ≪ 0.001 |
| CGPA vs. Depression | Pearson &  t-test | Weak correlation (r = –0.037),  small significance in t-test |
| Scholarship vs. Depression | t-test | Not statistically significant (p = 0.068), though visuals suggested a slight difference |

These tests confirmed that **sleep and family history** are reliable indicators of depression, whereas **academic performance and financial aid** are less definitive on their own.

**Machine Learning Applications**

We applied 3 models for each task:

**Regression (Predict PHQ-9 Score):**

* **Linear Regression**: Best performance (MSE = 0.1595, R² = 0.3422), showed academic pressure and financial stress as positive predictors.
* **KNN Regressor**: Weaker fit (R² = 0.237), poor generalization.
* **Decision Tree Regressor**: Interpretable but slightly overfit, performance similar to Linear (R² = 0.317).

**Classification (Depressed vs. Not Depressed):**

* **Logistic Regression**: Best performer (Accuracy = 78%, Recall = 0.84, ROC AUC = 0.845).
* **KNN Classifier**: Good recall but lower accuracy (75%).
* **Decision Tree**: Comparable to logistic, good interpretability, ROC AUC = 0.831.

**Selected Models**: Linear Regression for regression task, Logistic Regression for classification, due to balance of interpretability and accuracy.

**Tools & Technologies**

* **Python**, with libraries including **Pandas**, **Seaborn**, **Matplotlib**, **SciPy**, and **StatsModels**
* **Jupyter Notebook** for exploration
* **GitHub** for version control and progress tracking

**Limitations**

* **Data Imbalance**: Skew in sleep duration and depression status required careful consideration of metrics like recall and F1-score.
* **Self-Reported Data**: Bias in survey responses may affect reliability.
* **Single Time Point**: Longitudinal effects are not captured.
* **Generalization**: Results apply primarily to university populations and may not generalize beyond.

**Conclusion**

This project explored how sleep, stress, academic performance, and background factors relate to depression among students. Statistical testing and machine learning revealed that **sleep duration** and **family mental health history** are the strongest and most consistent predictors. While CGPA and financial aid show some associations, their explanatory power is weaker.

The combination of exploratory visualizations, hypothesis testing, and modeling allowed us to form a well-rounded understanding of depression predictors. Ultimately, our findings support targeted well-being interventions focusing on **sleep hygiene**, **stress management**, and **support for at-risk students** with known familial vulnerabilities.

**AI Assistance Disclosure**

Parts of this report—especially the structuring, drafting, and model interpretation—were developed with support from OpenAI’s ChatGPT. Specific support was received for understanding statistical tests, interpreting model results, and synthesizing insights across phases.