

Depression and Anxiety Predictor for High school in Kenya Data Report

1. Business Understanding

1.1 Business Overview

Mental Health issues among adolescents have become a growing concern in Kenya. The pressure of academics, social expectations and limited access to physiological support systems often lead to increased cases of depression and anxiety among high school students. This project aims at utilizing machine learning models and statistical analysis to build a predictive model that can identify students at risk of depression and anxiety based on their demographical, social and academic profiles.

1.2 Problem Statement

- *According to the Ministry of Health in 2021, up to 45% of secondary school students exhibited symptoms indicative of psychological distress. However, mental health screening remains largely absent in most schools. Even when mental health services are available, institutions may lack trained professionals and on top of that, widespread stigma surrounding mental illness often prevent students from seeking help at early stages.*
- *Right now when it comes to identifying teenagers struggling with depression or anxiety, schools and healthcare institutions are mostly relying on manual assessments by local counselors or healthcare workers - a process that is time consuming and not always reliable, and really hard to scale up across schools. As a result, many cases get missed, students start to fall behind in their studies, they start self medicating with drugs and, in some cases the risk of suicide actually increases.*

1.3 Objectives

Main Objectives

To develop a machine learning model capable of identifying depression and anxiety among Kenyan adolescents using responses from PHQ, GAD and other demographic assessments.

Specific Objectives:

1. To analyze adolescent survey data to find out what the major demographic factors mostly contribute to depression and anxiety.
2. To identify the major factors contributing to anxiety and/or depression.
3. To identify how different counties contribute to students' level of anxiety and depression.
4. To interpret model outputs and find out what are the most important factors that contribute to depression and anxiety prediction.

1.4 Research Questions

1. What demographic factors are linked to depression and anxiety in Kenyan teenagers?
2. Can PHQ and GAD scores be used to predict depression and anxiety levels?
3. How are students affected by depression and anxiety in different counties ?
4. What are the most important features contributing to depression and anxiety?

1.5 Success Criteria

- Accuracy:

At least 80% accuracy and balanced performance across precision, recall and F1-score.

- Reliability:

The model performs well on both training and test data, no overfitting.

- Interpretability:

Key features used in predictions (e.g. PHQ, GAD items, demographics) are clear and explainable to non-technical people.

- Ethical and Practical:

The system handles mental health data ethically and can be realistically implemented in school health systems.

- Impact:

Findings can inform early detection programs and resource allocation for adolescent mental health in Kenya.

1.6 Project Goals

- To build a machine learning model that classifies teenagers as depressed, anxious or none based on survey data.
- To combine demographic and psychological data (PHQ and GAD) for better mental health prediction.
- To get actionable insights that can help school administrators, counselors and policymakers support adolescent mental health.

2. Data Understanding

2.1 Data Description

- The dataset contains surveys from about 17,000 adolescents in different Kenyan counties. Each row contains a student's demographic background alongside the PHQ and GAD questions. Our main objective is to predict whether a person has Depression, Anxiety or none.
- For this project this dataset is suitable because it has many predictors:
 - Demographics: Age, Gender, Form, Religion, Boarding_day e.t.c
 - Depression Indicator Questions: PHQ'S
 - Anxiety Indicator Questions: GAD'S

2.2 Data Shape

The dataset shape is (17089, 34)

It contains columns such as

- participant_ID
- Age
- Gender
- Form

2.3 Data types

The data types are as follows : float64(29), object(5)

2.4 Initial Observations

Gender Distribution : Slightly more females than males.

Age Ranges : 13 - 19 Years.

Depression Rate 35-40%

Anxiety Rate : 30-33%

3 Data Preparation

3.1 Data Loading

Importing all necessary libraries to be used.

Loading the Dataset

3.1 Data Cleaning

- Handled Missing Values
- Removed any duplicates within the dataset
- Standardized categorical text entries
- Converted ordinal scales
- Handled all outliers.

3.2 Feature Engineering

1. Derived Variables
2. Encoding : Used One-Hot encoding for nominal categorical variables and used Label encoding for binary categories.
3. Scaling: Applied StandardScaler for continuous variables for models that are sensitive to scale.
4. VIF Analysis: Removed variables with VIF >10 to reduce multicollinearity

3.3.Data Split

The dataset was split into

- Training Set: 70%
- Test Set: 30%

Stratified Sampling was used to preserve class proportions

4 Exploratory Data Analysis

- Univariate Analysis:
 - The majority of participants have low to moderate PHQ-8 depression scores, indicating mild or minimal depressive symptoms in most of the sample. The distribution is positively skewed, with fewer individuals exhibiting high levels of depressive symptoms.
 - The mean item scores suggest that loss of interest and concentration difficulties are the most frequently reported depressive symptoms, while

psychomotor symptoms are least reported. This indicates **variability in symptom expression across the sample**.

- Bivariate Analysis
- ❖ Students in urban and high-performing counties (like Kiambu and Nairobi) particularly those in county and extracounty schools exhibit higher levels of depression and anxiety. These patterns may reflect the psychological costs of academic intensity, competitive environments, and limited rest or family interaction typical of such institutions.
- ❖ Meanwhile, rural counties (like Makueni and The analysis reveals notable differences in depression and anxiety prevalence across counties and school types in Kenya. Kiambu and Nairobi record the highest levels of both depression and anxiety, particularly in County and Extracounty schools, suggesting that students in more competitive or urbanized environments may experience greater psychological strain due to academic pressure and limited emotional support.
- ❖ In contrast, Makueni and Machakos show lower prevalence rates, possibly reflecting the protective influence of rural or community-based support systems. Overall, County and Extracounty schools appear more affected than Subcounty schools, indicating that school type and the associated boarding conditions, expectations, and competitiveness may play a significant role in shaping students' mental well-being. Machakos) report lower prevalence, suggesting that school environment and local context are key determinants of student wellbeing.

5. Modelling and Evaluation

The following models were tested :

1. Logistic Regression
2. Random Forest Classifier
3. XGBoost Classifier
4. Light GBM

5.2 Evaluation Matrix

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.81	0.79	0.77	0.78
Random Forest	0.86	0.84	0.82	0.83
XGBoost	0.88	0.86	0.85	0.85

5.3 Best Model

XGBoost Classifier emerged as the best model with the highest accuracy and has a balanced precision and recall. This made it suitable for deployment in sensitive screening contexts.

The models focused on **predicting people with anxiety and depression**, however it does not just predict whether a person is depressed or not rather *it predicts the severity of the depression or anxiety*. For this project **recall** was used as the major metric. All the models used for this project performed well while identifying the various depression and anxiety levels.

- **Logistic Regression**- The model *works very well for the common classes (No depression and Mild)* which means it does not miss many of these cases. It however seems to *struggle more with the higher severity levels; categories 3 and 4*, where **recall is very low**. In actual applications, this means that the model may be underestimating the cases of serious depression.
- **Random Forest**- The Random Forest model shows *acceptable overall accuracy and good recall for mild and moderate levels*, which suggests it can identify symptoms early. However, it *underperformed when identifying cases of severe depression and anxiety*, which is the most important level of mental health diagnosis to identify in practice. A *recall of 0.50 on severe depression cases and 0.583 recall for severe anxiety* indicated that class balancing in training data may help identify individuals in severe cases of mental health diagnoses.
- **XGBoost**- In comparison to Random Forest, the adjusted XGBoost model shows *improved and more consistently high recall rates, but primarily for moderate-risk cases*. The model still **struggled with severe depression recall (40%)**, it fared better in identifying patients with severe anxiety (recall = 0.699). Given the XGBoost model had a *higher overall recall and accuracy*, it is seen as a *better model for identification of mild and moderate cases*, and slightly better in the severe case category.
- **LightGBM** - LightGBM **did well overall for depression with an overall accuracy of 83.8%. High recall for the mild and moderate classes** (0, 1, and 2), suggesting that the model will get people on the lower end of the severity of depression spectrum. *Recall was significantly lower for the more severe classes (3, and 4)*, with recall reported as 0.624 for class 3 and 0.545 for class 4. This shows there are instances where the *model does not identify someone with severe or very severe depression*, and this may be related to fewer cases of severe depression in the dataset. Therefore, even with a high overall accuracy, recall for the more severe levels is lower, which shows the *model is not as sensitive in the more serious classes, which is a considerable limitation when detecting individuals that may need the most help*. For anxiety, the overall accuracy was a little higher, at 88%. **The recall values were also fairly good overall for anxiety**, with the mild and moderate classes being above 0.80, and the severe class (3) being at 0.798. Therefore, this indicates that the *model is doing a better job detecting anxiety than depression* and was able to catch most levels of anxiety correctly. However, again, the model *missed*

detecting some of the very severe cases, but was not missing cases as well when gauging depression.

6. Insights and Interpretation

6.1 Key predictors of Depression and Anxiety

1. Low Family Support
2. High Academic Stress
3. Poor Sleep Quality
4. Low Self-esteem

6.2 Interpretation

The model highlights that physiological factors are stronger predictors than purely academic and demographic factors.

These findings align with global studies which emphasize environmental and emotional well being as major determinants of adolescent mental health.

6.3 Visualization Highlights

- Bar charts showing depression rates by gender and grade
- Heatmap of correlation among top variables
- Feature Importance plot from XGBoost showing relative contributions of predictors

7 Limitations

- Self-reported data may be biased or inaccurate
- Dataset is cross sectional ie cannot infer casualty
- Limited geographical coverage ie limited regions in Kenya
- Class imbalance slightly affected recall in class minority

8. Summary and Recommendations

8.1 Summary

The project successfully developed a robust predictive model using student survey data.

- The XGBoost classifier achieved an accuracy of 88%
- Key determinants of depression and anxiety are identified.
- The model provides a data driven basis for early intervention in schools.
- When we trained our models, we processed the original dataset through a preprocessing pipeline that normalized numeric features, then compressed numeric features, and used one-hot encoding. Therefore, when the models were built, they were not able to identify the original categorical names such as Gender or School_Type but used technical naming conventions such as cat__School_Type_Extracounty or num__age. The use of the code was to reverse the transformation of the feature names, so when looking at summary or SHAP plots, we could remember what inputs contributed to predictions by the models.

- The summary above helps us understand what contributed to a model making a certain prediction.

8.2 Recommendations

1. Deploy Pilot models in schools to flag high risk students confidentially
2. Enhance counselling services and awareness programs targeting family and peer support.
3. Regular mental health screenings should be mandatory in schools.
4. Future studies should collect longitudinal data for casual inference
5. Developing an interactive dashboard for policy visualization.

9. Deployment

A simple streamlit web application ([app.py](#)) was built to allow users (Teachers and Counsellors) to input student data and will get real time prediction.

The app loads the trained model(LightGBM_model.pkl) and displays the predicted risk category.

Performance varies by severity level and type of symptoms, but by deploying all four models and using the best output for each user will add reliability. Overall, this is more of an ensemble-style approach that ensures the results are not driven by failure of accuracy from a single model.

- Deploy this model as an early mental health screening tool in Kenyan schools.
- Collect more data from all geographical locations in Kenya.
- The tool should not be used for diagnosis.

10. Conclusion

- Machine learning can effectively identify students at potential risk of mental health issues.
- Key predictors include sleep quality, academic pressure and family relationships.
- The final model achieved an accuracy of 88% and can support early mental health screening.