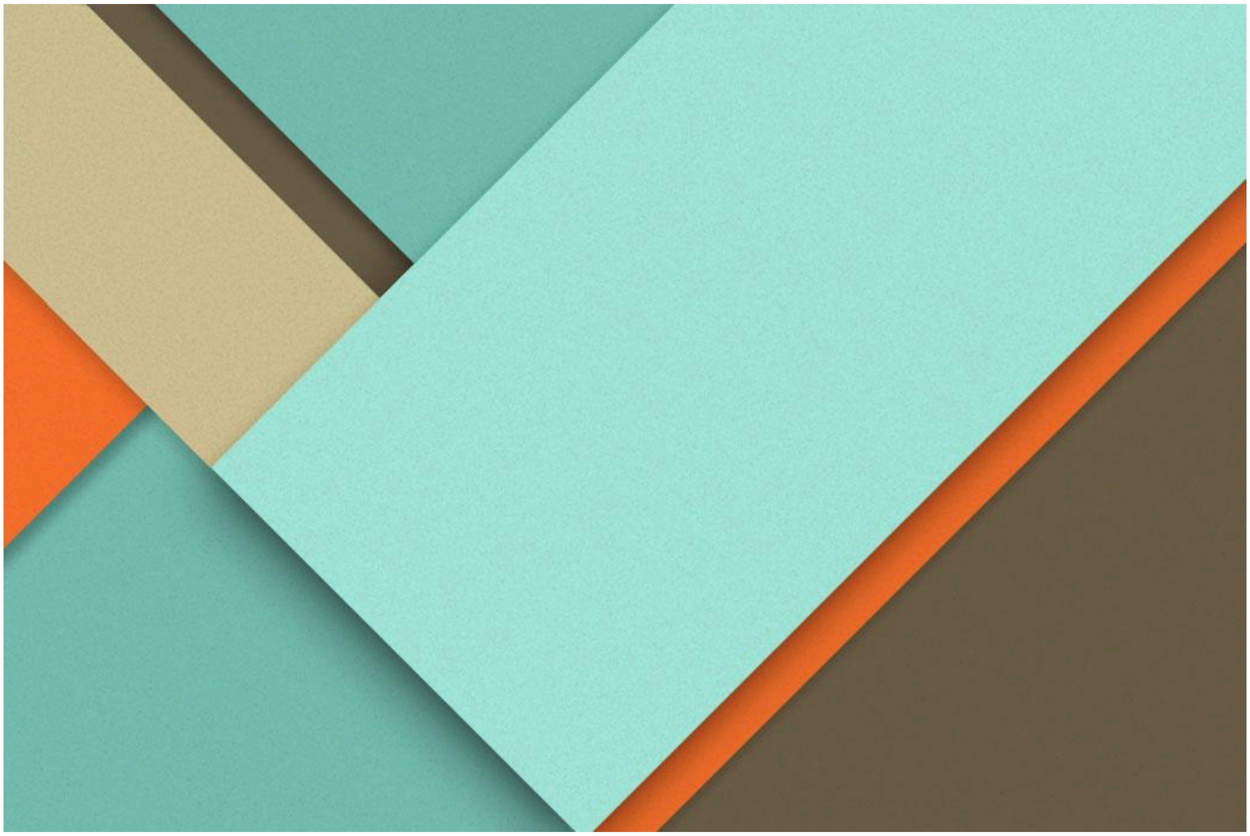


Exploring the Relationship Between Social Media Usage and Anxiety: Part 2

07.02.2026

—

Hayley Lee

Overview

For this project, a dataset containing individuals' social media usage patterns and clinically reported anxiety levels will be used to support a machine learning classification task. The objective is to train and evaluate a model that can predict a user's anxiety severity based on features derived from social media behavior.

Goals

1. Select suitable machine learning (ML) models for this project.
2. Perform model training, validation and testing.
3. Evaluate and compare model performance, while implementing performance improvements where applicable.
4. For the conclusion, explain current results and suggestions for future performance improvements and areas of study relating to the project.

Model Selection

When analysing the dependent variable (Anxiety Severity), the dataset in general and the hypothesis being tested, the following observations can be made:

- GAD_7_Severity is split into multiple, distinct categories of Mild, Minimal, Moderate and Severe.
- The dataset in general is highly structured, but not large in size with 8000 unique users.

The dataset has highly structured and labelled input and output data, making it suitable for predicting a specified output using supervised learning. Since the dependent variable is split into multiple, distinct categories, the model should be capable of multiclass classification. Classification models such as Logistic Regression and Support Vector Machine (SVM) would not be suitable without one-versus-one or one-versus-all binary transformations (Keita, 2024) , as they are primarily suited for binary classification problems (only two classes to predict) (MathsWorks, 2026). Other models which support multiclass classification include: Decision Trees, Random Forests, Naive Bayes, XG Boost and Neural Networks (Kuo, 2022; Pirge, 2020).

This report will evaluate and compare the effectiveness of Decision Trees (DT), Random Forests (RF) and XG Boost (XGB) models in predicting the anxiety severity class of users based on their social media habits.

Model Training, Validation and Testing

Preprocessing

The steps included for data preprocessing included:

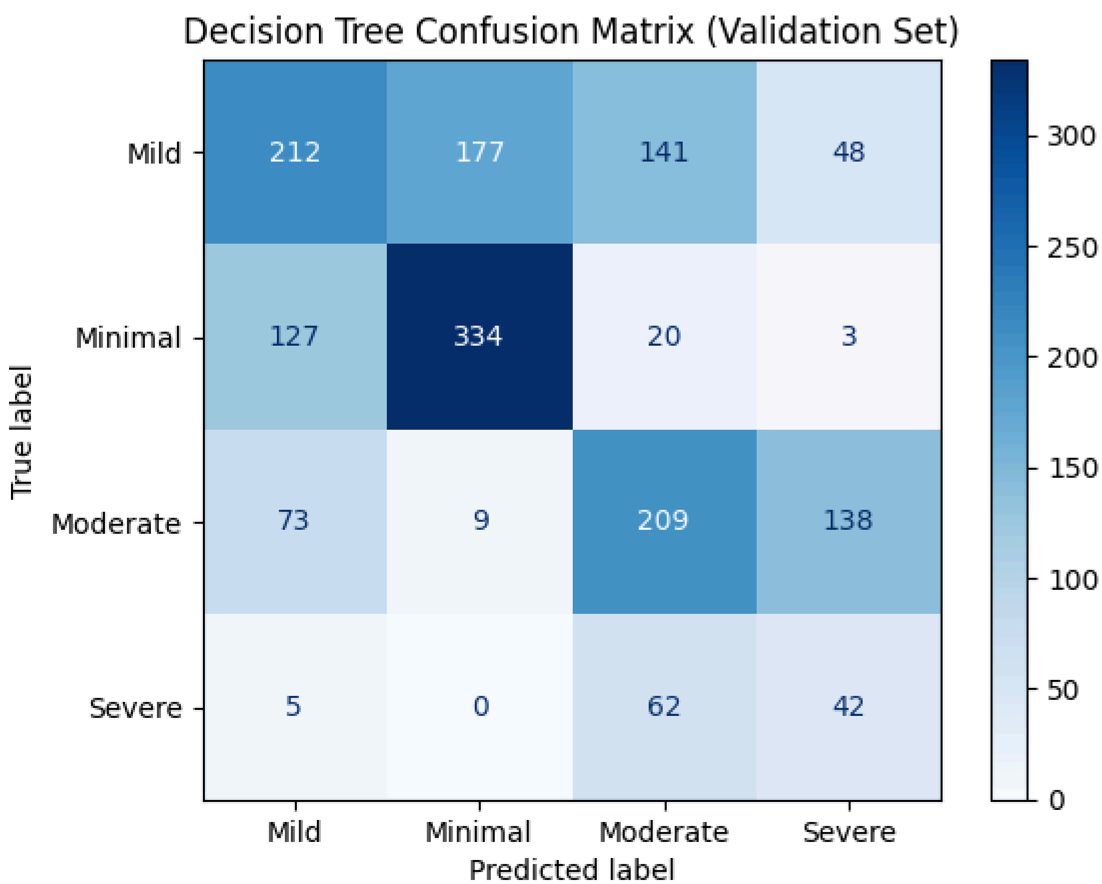
- Data Stratification of Anxiety Score: Used to minimise the effect of imbalance in severe anxiety cases (previously found to make up only 6.8% of the data). This helped minimise bias introduced by the lack of severe anxiety cases (GeeksforGeeks, 2025a).
- Label encoding: Converting categorical text data to numerical inputs that the model can understand (GeeksforGeeks, 2025b).
- Specific anxiety evaluation scoring was removed, to avoid the model using this to directly infer a users anxiety category without social media habits.
- Depression history was removed to determine if social media habits alone would impact a user's anxiety level.

Training, Validation and Testing

For code used, see jupyter notebook in github.

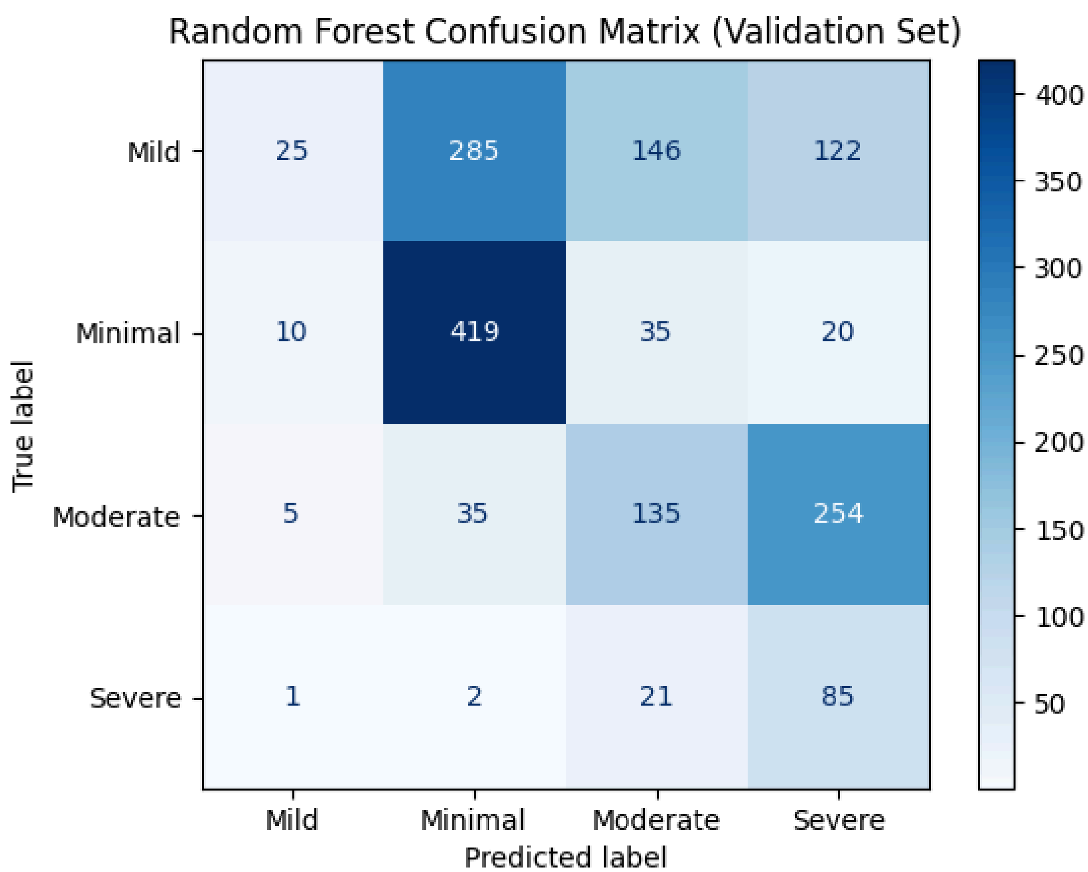
Initial model scores and confusion matrices for DT, RF and XGB without hyperparameter tuning:

```
Decision Tree
              precision    recall  f1-score   support

        Mild       0.51      0.37      0.43       578
     Minimal       0.64      0.69      0.67       484
    Moderate       0.48      0.49      0.49       429
      Severe       0.18      0.39      0.25       109

    accuracy                           0.50      1600
   macro avg       0.45      0.48      0.46      1600
weighted avg       0.52      0.50      0.50      1600
```



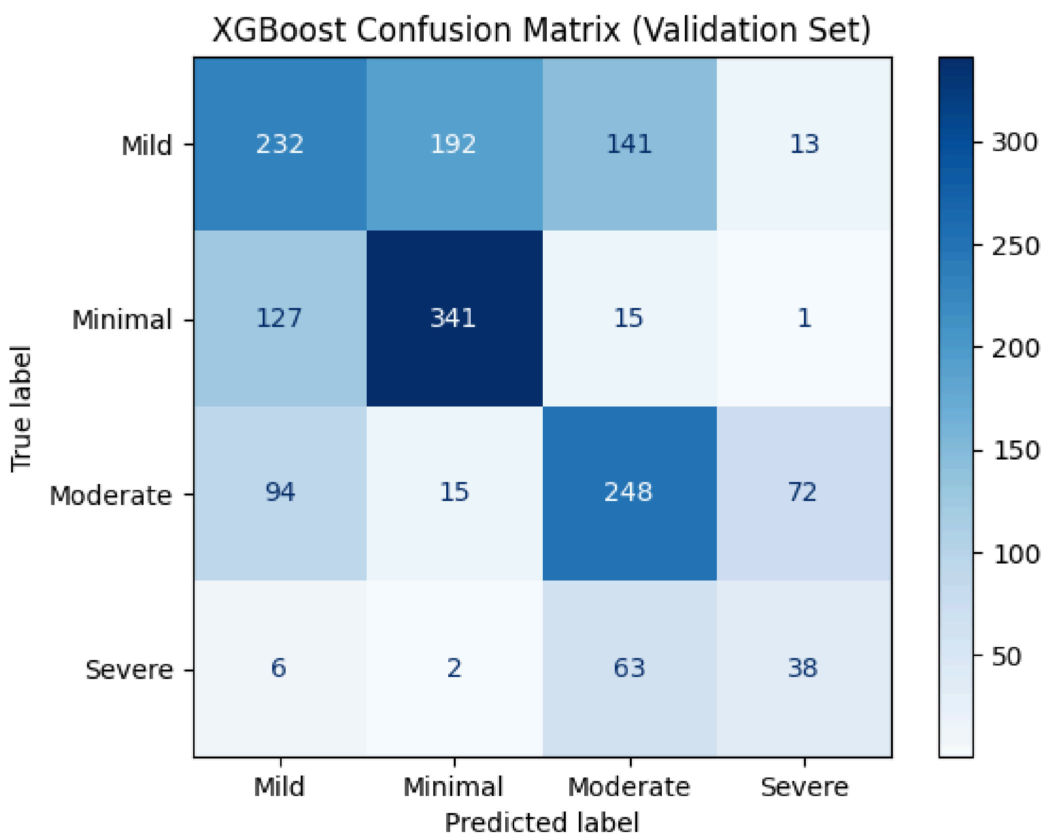
Random Forest

	precision	recall	f1-score	support
Mild	0.61	0.04	0.08	578
Minimal	0.57	0.87	0.68	484
Moderate	0.40	0.31	0.35	429
Severe	0.18	0.78	0.29	109
accuracy			0.41	1600
macro avg	0.44	0.50	0.35	1600
weighted avg	0.51	0.41	0.35	1600



XGBoost

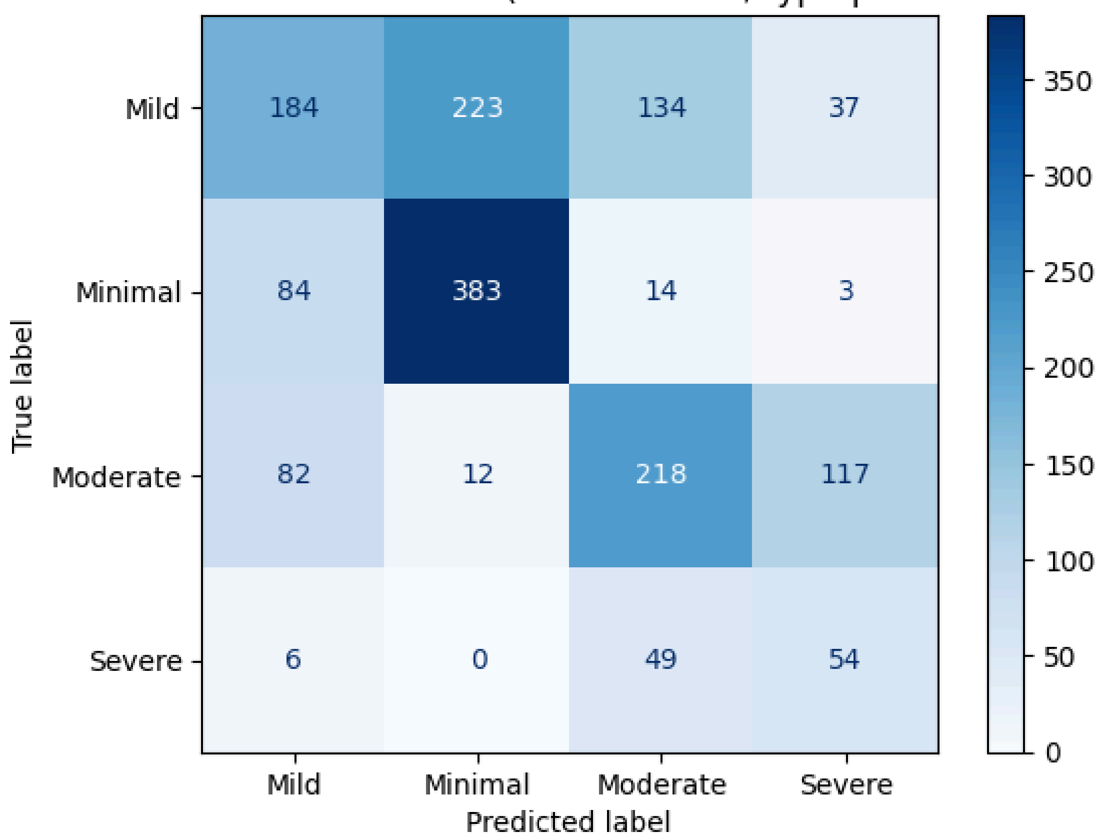
	precision	recall	f1-score	support
Mild	0.51	0.40	0.45	578
Minimal	0.62	0.70	0.66	484
Moderate	0.53	0.58	0.55	429
Severe	0.31	0.35	0.33	109
accuracy			0.54	1600
macro avg	0.49	0.51	0.50	1600
weighted avg	0.53	0.54	0.53	1600



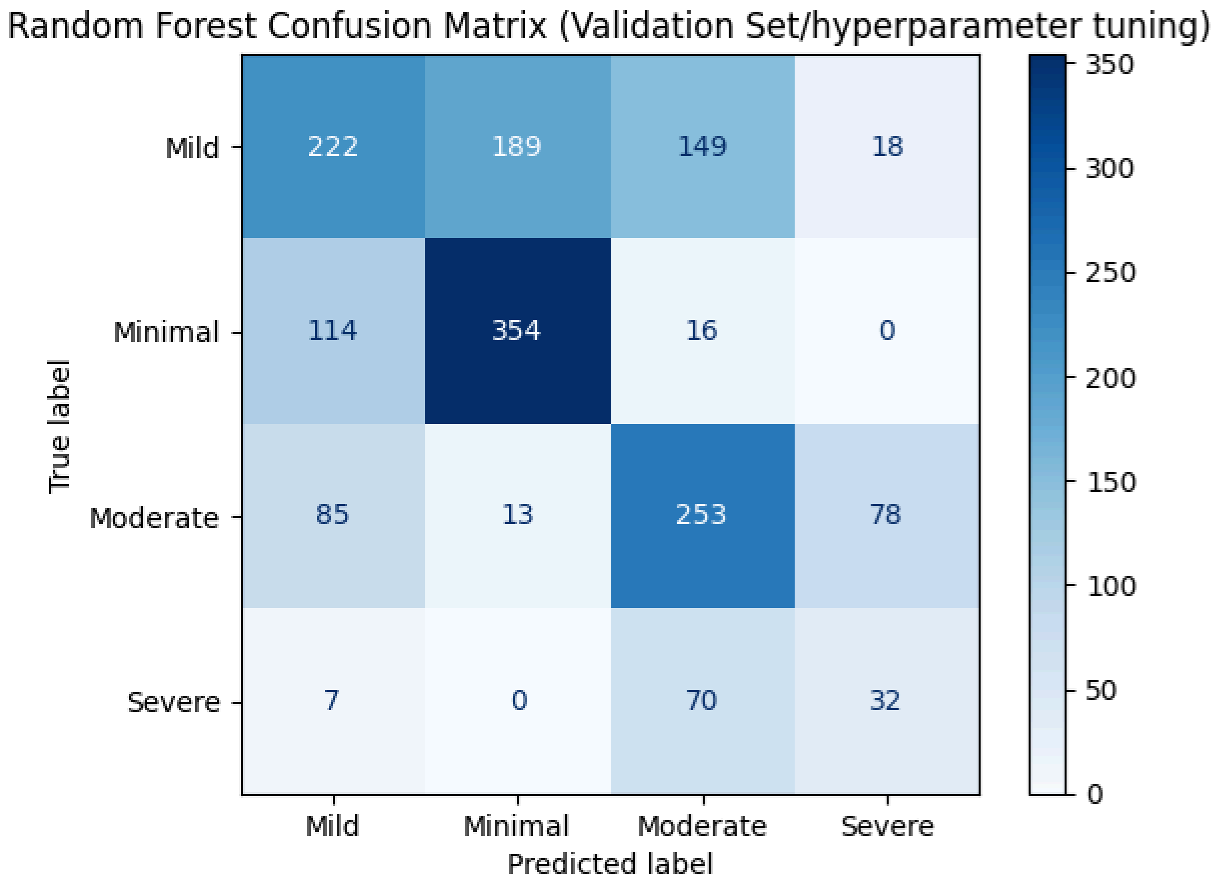
Model scores and confusion matrices for DT, RF and XGB, after hyperparameter tuning:

```
Best Decision Tree CV F1: 0.5310749695444835
              precision    recall  f1-score   support

        Mild       0.52      0.32      0.39       578
     Minimal       0.62      0.79      0.70       484
    Moderate       0.53      0.51      0.52       429
      Severe       0.26      0.50      0.34       109

    accuracy                           0.52      1600
   macro avg       0.48      0.53      0.49      1600
weighted avg       0.53      0.52      0.51      1600
```

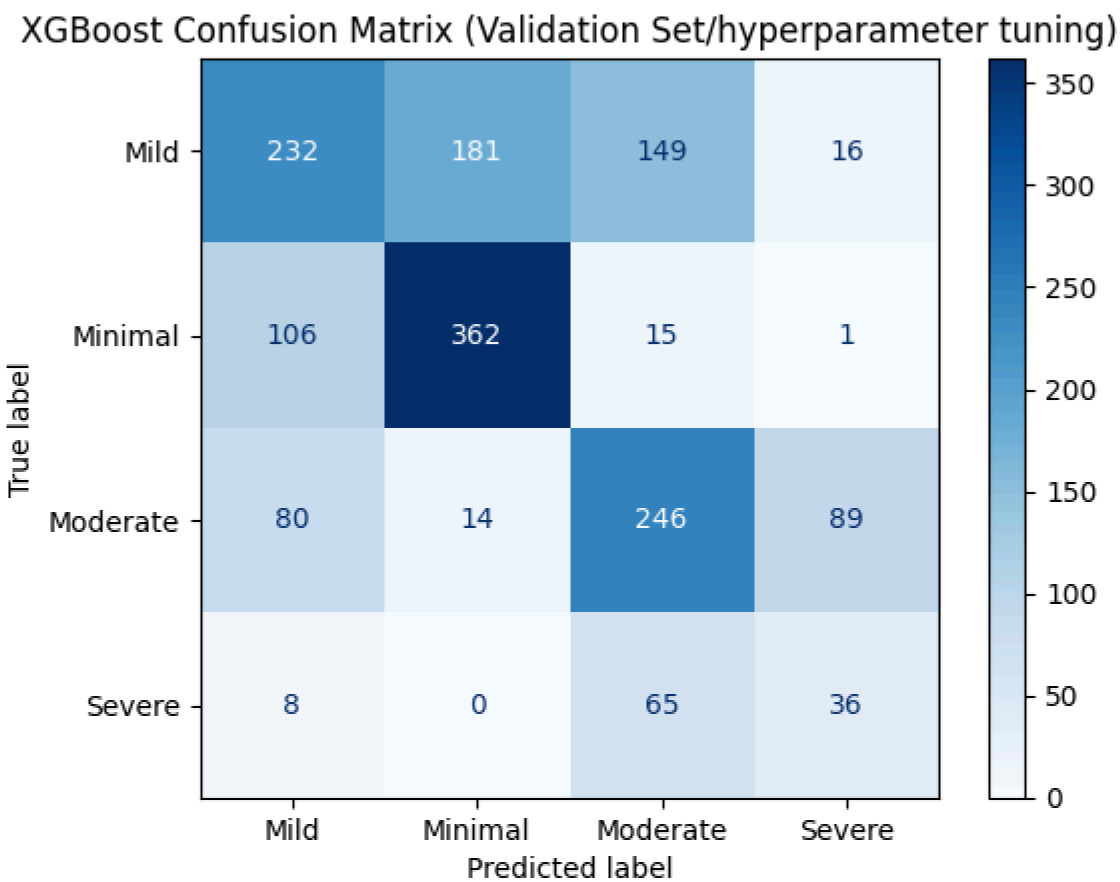
Decision Tree Confusion Matrix (Validation Set/hyperparameter tuning)



```
Best Random Forest CV F1: 0.5569697331520742
              precision    recall  f1-score   support

        Mild       0.52      0.38      0.44       578
     Minimal       0.64      0.73      0.68       484
    Moderate       0.52      0.59      0.55       429
      Severe       0.25      0.29      0.27       109

    accuracy                           0.54      1600
   macro avg       0.48      0.50      0.49      1600
weighted avg       0.54      0.54      0.53      1600
```



```
Best XGBoost CV F1: 0.5563803697952371
              precision    recall  f1-score   support

        Mild       0.54      0.40      0.46       578
     Minimal       0.65      0.75      0.70       484
    Moderate       0.52      0.57      0.54       429
      Severe       0.25      0.33      0.29       109

    accuracy                           0.55      1600
   macro avg       0.49      0.51      0.50      1600
weighted avg       0.55      0.55      0.54      1600
```



When comparing the results between the untuned/tuned models, the following observations can be made:

Decision Tree

- Accuracy increased from 0.50 to 0.52
- Macro Average F1 increased from 0.46 to 0.49
- Weighted Average F1 increased from 0.50 to 0.51
- There was an increase in recall for Minimal from 0.69 to 0.79
- Moderate F1 increased from 0.49 to 0.52
- Severe recall improved but precision still low

In general, the model improved predictions across all classes, evidenced by the macro average F1 score increase. The model shows bias towards the Minimal class.

Random Forest

- Accuracy increased from 0.41 to 0.54
- Macro Average F1 increased from 0.35 - 0.49
- Weighted F1 increased from 0.35 to 0.53
- Before tuning, Mild recall was only 0.04. This was improved to 0.38

The class Mild was previously almost ignored, with an extremely low recall. This was balanced through tuning. This model over-predicted Severe, with a recall of 0.78 but precision of 0.18.

XGBoost

- Accuracy increased from 0.54 to 0.55
- Macro Average F1 remained the same
- Weighted F1 increased slightly from 0.53 to 0.54
- Minimal was predicted the best with an F1 score of 0.70
- Severe prediction worsens

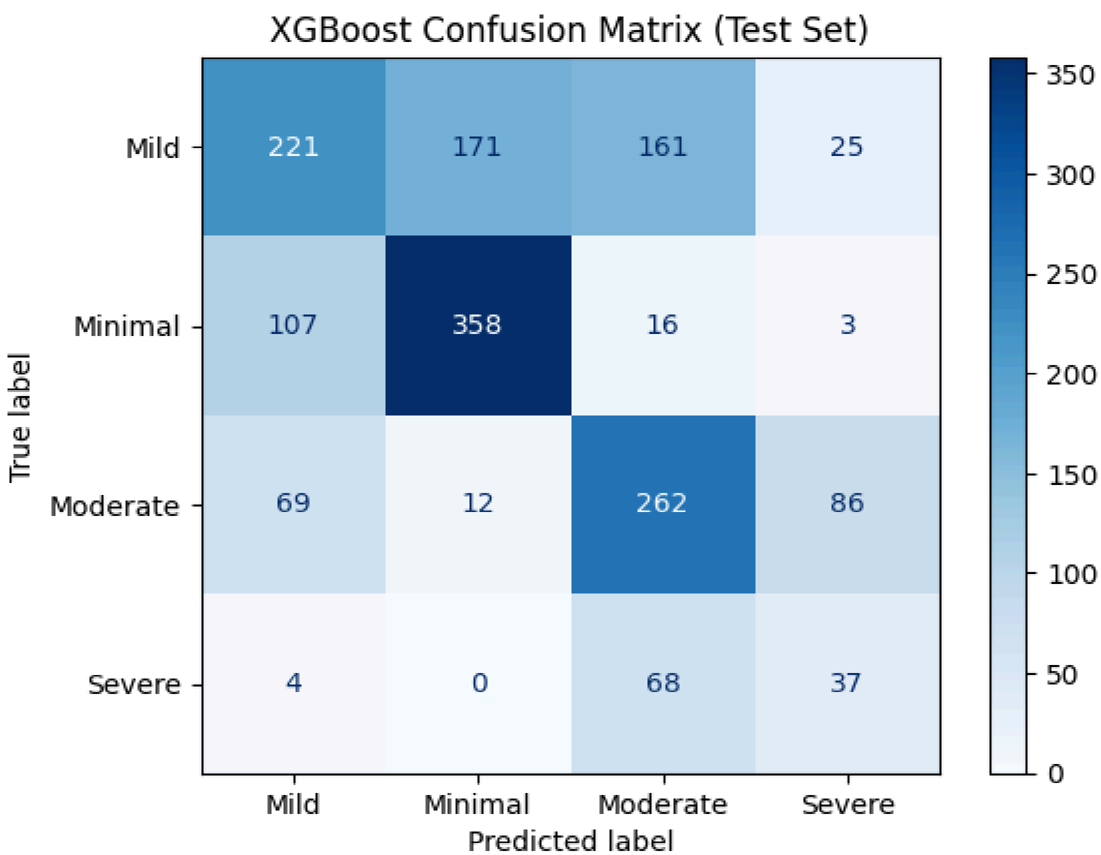
XGBoost results didn’t improve much, but gave consistent results without affecting balance.

Final Model Selection

After hyperparameter tuning, all models showed improved performance. Random Forest improved the most, correcting Severe class imbalance issues present before optimisation. However, tuned XGBoost achieved the best overall performance, with the highest accuracy (0.55), weighted F1-score (0.54), and balanced class-wise metrics. Therefore, XGBoost was selected as the final model.

Final results using tuned XGBoost and test set data:

```
XGBoost
              precision    recall  f1-score   support

        Mild       0.55      0.38      0.45       578
     Minimal       0.66      0.74      0.70       484
    Moderate       0.52      0.61      0.56       429
      Severe       0.25      0.34      0.28       109

    accuracy                           0.55      1600
   macro avg       0.49      0.52      0.50      1600
weighted avg       0.55      0.55      0.54      1600
```



From the above results, the following observations can be made:

- Minimal was the best predicted class with a F1 score of 0.74
- Moderate class identification was fair, with an F1 score of 0.56. The model is slightly more biased towards Moderate due to 0.61 recall.
- Mild and Severe case identification still show reduced performance after hyperparameter tuning.

Conclusion

In conclusion, three models suitable for multiclass classification problems were selected and trained to predict the anxiety level of users, using their social media habits. DT, RF and XGB models were selected for their ability to predict multiclass results without needing to incorporate further one-versus-one or one-versus-all binary transformations and their suitability towards highly structured data. During model validation, hyperparameter tuning was used to improve the performance of the chosen models. XGB was found to be the best performer, as it handled the class imbalance of severity levels the best (shown by the macro average F1 score being the highest).

When the tuned XGB model was used on unseen data (test set), it showed strong performance for all classes except Severe, despite data stratification and class weight balancing being used. Previous data exploration (Figure 1) shows the dataset significantly lacked Severe cases, which would affect the model in this way.

```
GAD_7_Severity
Mild        36.1375
Minimal     30.2375
Moderate    26.8125
Severe       6.8125
Name: proportion, dtype: float64
```

Figure 1: The class distribution of the dataset before splitting, given in percentages.

Since the Minimal class was approximately one third of the data, there were sufficient examples for the model to use during training for reliable results. This is shown by Minimal having the highest recall and F1 score. Despite Mild being the most represented class, it did not perform as well as Minimal or Moderate, which suggests there was confusion when the model found overlapping features between Mild and its neighbouring classes (Minimal and Moderate).

Suggestions for future research and improvements include:

- Since this research focussed solely on the effect of social media habits on mental health, a user's history of depression was omitted. This could be included in future experiments, to see if a history of depression makes a user more susceptible to higher anxiety levels when using social media.

- Different ML models suitable for multiclass classification (e.g. Naive Bayes, Neural Networks) can be tested and compared to XGB, in case there is a better performing alternative.
- Collect more Severe anxiety samples to balance out the dataset.
- Use synthetic oversampling methods such as SMOTE (Synthetic Minority Oversampling Technique). This method generates synthetic data for the minority class (in this case Severe) using interpolation of existing nearest neighbour data points (GeeksforGeeks, 2025c). This is better than simply duplicating data entries.

References

GeeksforGeeks (2025a) Stratified sampling in machine learning, GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/stratified-sampling-in-machine-learning/ (Accessed: 07 February 2026).

GeeksforGeeks (2025b) Categorical data encoding techniques in machine learning, GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/categorical-data-encoding-techniques-in-machine-learning/ (Accessed: 07 February 2026).

GeeksforGeeks (2025b) Smote for imbalanced classification with python, GeeksforGeeks. Available at: https://www.geeksforgeeks.org/machine-learning/smote-for-imbalanced-classification-with-python/ (Accessed: 08 February 2026).

Keita, Z. (2024) Classification in machine learning: A guide for beginners | datacamp, DataCamp. Available at: https://www.datacamp.com/blog/classification-machine-learning (Accessed: 07 February 2026).

Kuo, C. (2022) A wide variety of models for multi-class classification | by Chris Kuo/dr. Dataman | Dataman in AI | Medium, Medium. Available at: https://medium.com/dataman-in-ai/a-wide-choice-for-modeling-multi-class-classifications-d97073ff4ec8 (Accessed: 07 February 2026).

MathsWorks (2026) Choosing the Best Machine Learning Classification model and avoiding overfitting, MATLAB & Simulink. Available at: https://www.mathworks.com/campaigns/offers/next/choosing-the-best-machine-learning-classification-model-and-avoiding-overfitting.html (Accessed: 07 February 2026).

Pirge, G. (2020) Performance comparison of multi-class classification algorithms | by Gursev Pirge | Medium, Medium. Available at: https://gursev--pirge-medium-com.translate.goog/performance-comparison-of-multi-class-

classification-algorithms-606e8ba4e0ee?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=tc (Accessed: 07 February 2026).