Kareem Hasan
INST414

Data Science Lifecycle Option

Executive Summary for Non-Technical Stakeholders

This project shows how we can use machine learning to identify people at higher risk of serious mood swings which is one of many signs for conditions like anxiety or depression. Using survey data from Bangladesh, we trained and compared tree-based models to predict mood swings on factors like stress, age, and occupation. The project's results can inform NGO's, nonprofits and mental health organizations to target early interventions and outreach to improve mental health outcomes.

Problem Statement and Business/Social Context of this Work.

Mood swings, specifically defined as rapid and intense fluctuations in emotional states, pose significant challenges to mental health and overall well-being. In Bangladesh, where the RHMCD-20 dataset comes from, is the data that I'll be using for analysis. In Bangladeshi culture there are stigmas surrounding mental health that often lead to under reporting, and inadequate treatment of conditions like depression/anxiety. There is a lack of awareness, making it difficult for individuals to seek professional help. This gap in mental health care makes it crucial to actually study predictive factors that can contribute to mood instability and more. The RHMCD-20 dataset provides a wide collection of variables to look at, including things like age, gender, duration indoors, habit changes, mental health history, weight change, and coping struggles. This dataset gives us the opportunity to analyze the potential contributing factors of mood swings within the Bangladeshi population. Not just that, but understanding these relationships can help in finding at-risk individuals and identifying interventions earlier.

This is a problem worth tackling for many reasons. Mental health issues in general can adversely affect personal relationships, job performance, and overall quality of life. By analyzing the dataset, we can identify patterns and signs associated with mood instability, or potentially find predictors early on and implement support. This is a pressing issue on its own, and isn't reported enough. In addition, the long term psychological effects of COVID-19 are enough of a reason this should be researched more. Studies have shown that extended periods of quarantine or social restrictions have significantly influenced mental health. Research by Vindegaard & Benros in The Lancet Psychiatry further confirms that pandemic-related mental health issues have increased like anxiety, depression, and mood disorders (Vindegaard & Benros, 2020). By analyzing our dataset, we can identify behavioral trends that may indicate early signs of mood instability and suggest strategies for improved mental health interventions.

In addition, this research can provide insight into the broader implications of mental health prediction using machine learning models. Many traditional diagnostic tools rely on self-reported data, which can be subject to bias or inaccuracies. By leveraging decision trees for predictive modeling, we can create a data-driven approach that may serve as an additional tool for mental health professionals. This research could also inform policies aimed at increasing mental health awareness and accessibility to treatment in regions where mental health care is underdeveloped.

Addressing mood swings as a predictor of underlying mental health conditions is important for early detection and intervention. If left unaddressed, prolonged mood instability can develop into more severe conditions such as major depressive disorder (MDD) or bipolar disorder. The impact of undiagnosed and untreated mood disorders extends beyond the individual, affecting families, workplaces, and even national productivity (Vindegaard & Benros, 2020). As mental health becomes an increasingly recognized public health concern, it is critical to utilize technological advancements in data science to support early detection and treatment strategies.

Understanding and predicting mood swings is also great for NGOs, public health organizations, and policymakers. We can identify at-risk individuals earlier and allocate resources to improve outcomes at both personal and societal levels. We can also contribute to a growing body of research that aims to make mental health care more proactive rather than reactive. This research is urgent as the long-term effects of the COVID-19 pandemic continue to seep/trickle in and disrupt well-being across the globe.

Methodology

There have been many more recent studies that have actually employed machine learning techniques to predict depressive symptoms without relying on traditional survey questions. For example, research from a National Library of Medicine article discussed a model that used demographic as well as behavioral data to identify individuals at risk of depression (Amin, 2024). This showed the potential of not having to rely on self-report for these variables. In addition, the same dataset has also been used to investigate mental health depression during quarantine life using machine learning techniques. These recent/relevant studies, using machine learning algorithms and other modeling techniques, have shown how important it is to keep doing these studies on mental health.

Another approach has been the use of digital tracing to assess mood swings. By analyzing social behaviors, digital footprints, and self-reported experiences, researchers have been developing predictive models for mood disorders. There are many datasets like RHMCD-20 that provide us an opportunity to replicate these methods in more places and not just Bangladesh (Pelt et al., 2024).

Kareem Hasan
INST414

Data Preparation

The initial phase will involve cleaning the dataset to address any inconsistencies or missing values. For data encoding, we applied ordinal encoding to variables with a natural order, like age and days spent indoors, and binary encoding for gender and coping struggles. Occupation, being a nominal variable, was one-hot encoded g to prevent artificial relationships. Yes/No/Maybe variables like growing stress and weight change were also ordinally encoded. This step is crucial to prepare the data for more accurate analyses. We do this because it prevents variables with larger scales from influencing the models decisions and improves algorithm performance.

‘‘‘

```
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split, cross_val_score
from xgboost import XGBRegressor
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import mean_absolute_error
from catboost import CatBoostRegressor

df = pd.read_csv("sprint1data.csv")
X = df[["Growing_Stress", "Weight_Change", "Social_Weakness"]]
y = df["Mood_Swings"]
encoder = LabelEncoder()
X_encoded = X.copy()
for col in X_encoded.columns:
    X_encoded[col] = encoder.fit_transform(X_encoded[col])
y_encoded = encoder.fit_transform(y)

X_train, X_test, y_train, y_test = train_test_split(
    X_encoded, y_encoded, test_size=0.2, random_state=42
)
```

’’’

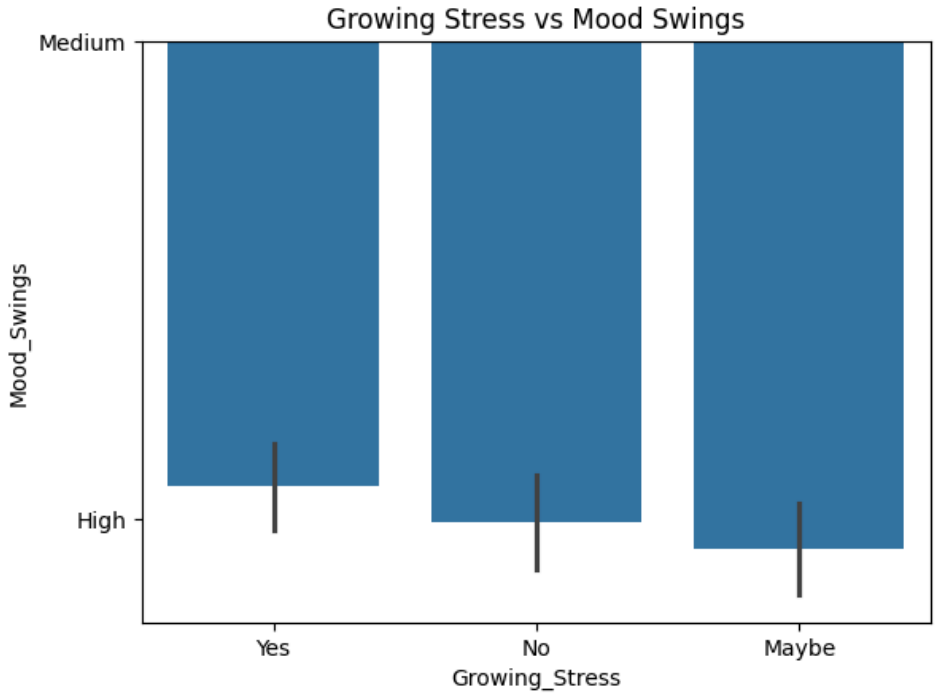
Dataset

The primary data source for this project is the RHMCD-20 dataset, which offers lots of information on various different mental health factors. The dataset is pretty comprehensive and I ensure it is sufficiently large enough for predictive modeling/analytics. The documentation for the dataset is in the Mendeley Data article (Amin & Salehin, 2024)

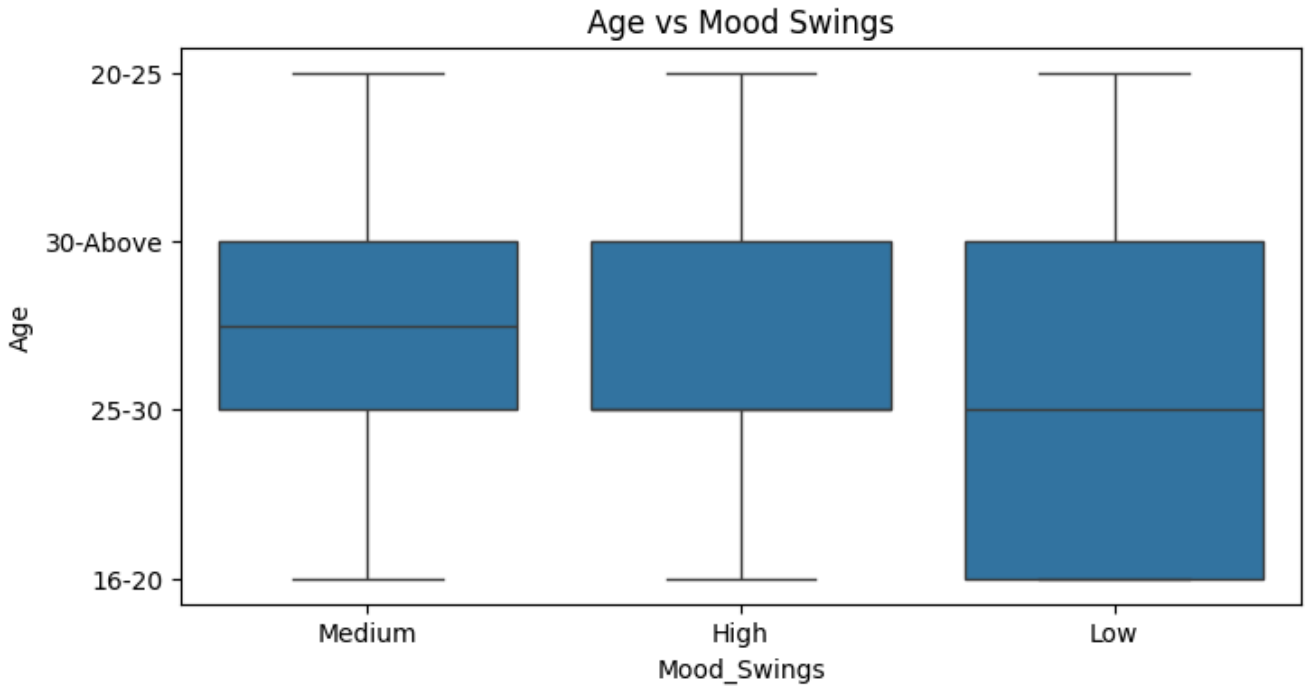
Feature Engineering

Feature engineering focused on constructing composite features like a “Stress Index” and a “Mental Health Vulnerability” score, and even interaction terms like Age-Isolation to show possible effects of quarantine on different age groups.

- Stress Index: Mean of Growing_Stress_encoded, Quarantine_Frustrations_encoded, and Coping_Struggles_encoded.
    - Justification: Creates a composite measure of overall stress levels experienced during the pandemic.
- Mental Health Vulnerability: Mean of Mental_Health_History_encoded, Mood_Swings_encoded, and Social_Weakness_encoded.
    - Justification: Combines factors that may indicate vulnerability to mental health challenges.
- Age-Isolation Interaction: Product of Age_encoded and Days_Indoors_encoded.
    - Justification: Captures the interaction effect between age and isolation, which may affect mental health differently across age groups.
- Quarantine Impact: Product of Days_Indoors_encoded and Quarantine_Frustrations_encoded.
    - Justification: Measures the combined effect of isolation duration and frustration level.
- High Risk Mental Health: Binary feature indicating high-risk individuals based on Mental_Health_History, Mood_Swings, and Coping_Struggles.
    - Justification: Identifies individuals who may be at elevated risk for mental health challenges.
- Routine Disruption: Product of Days_Indoors_encoded and Changes_Habits_encoded.
    - Justification: Measures the level of disruption to daily routines caused by pandemic restrictions.



Kareem Hasan
INST414



Modeling approaches

This project uses a Decision Tree Regressor to analyze the relationship between various demographic and psychological factors and coping struggles. Decision trees are effective for predictive modeling because they can handle both categorical and numerical data. The dataset contains multiple variables such as age, gender, occupation, time spent indoors, stress levels, changes in habits, and more. Our objective is to predict individuals' mood swings based on these factors. The Decision Tree model was chosen for its simplicity in identifying patterns within the dataset. It breaks the data into smaller subsets based on feature importance, leading to a set of if-then decision rules that run a prediction. The model is also well-suited for handling psychological and behavioral data.

To successfully execute this project, you will need essential technical skills like data preprocessing and exploratory data analysis. It's important to have proficiency in handling missing data, handling categorical variables, and being able to identify correlations. In addition to pandas, we implemented a model using Scikit-learn, where it's essential to know tuning for the machine learning algorithm. Overfitting happens when a model learns noise in the training data as if it were a true pattern, leading it to perform well on training data but bad on new, unseen data. Techniques like pruning help reduce overfitting by removing branches of the tree that add low predictive power. Hyperparameters for XGBoost (like learning rate, max depth, and subsample ratio) and CatBoost (iterations, depth, and learning rate) were set and further tuning will be done in the future using grid search.

XGBoost model

```
xgb_model = XGBRegressor(
    n_estimators=100,
    learning_rate=0.1,
    max_depth=3,
    subsample=0.8,
    colsample_bytree=0.8,
    objective='reg:squarederror',
    random_state=37
)
xgb_model.fit(X_train, y_train)
xgb_pred = xgb_model.predict(X_test)
xgb_mae = mean_absolute_error(y_test, xgb_pred)
print("XGBoost:")
print("MAE:", xgb_mae)

xgb_cv_scores = cross_val_score(
    xgb_model, X_encoded, y_encoded,
    cv=10, scoring='neg_mean_absolute_error'
)
print("Cross-validation MAE scores:", -xgb_cv_scores)
print("Average CV MAE:", -xgb_cv_scores.mean())

# CatBoost model
cat_model = CatBoostRegressor(
    iterations=100,
    learning_rate=0.1,
    depth=3,
    loss_function='RMSE',
    random_seed=37,
    verbose=0
)
cat_model.fit(X_train, y_train)
cat_pred = cat_model.predict(X_test)
cat_mae = mean_absolute_error(y_test, cat_pred)
print("CatBoost:")
print("Mean Absolute Error:", cat_mae)

cat_cv_scores = cross_val_score(
    cat_model, X_encoded, y_encoded,
    cv=10, scoring='neg_mean_absolute_error'
```

```
)
print("Cross-validation MAE scores:", -cat_cv_scores)
print("Average CV MAE:", -cat_cv_scores.mean())
```

Code Repository:
https://github.com/kareemhasan27/414_Project.git

Model Evaluation and Selection

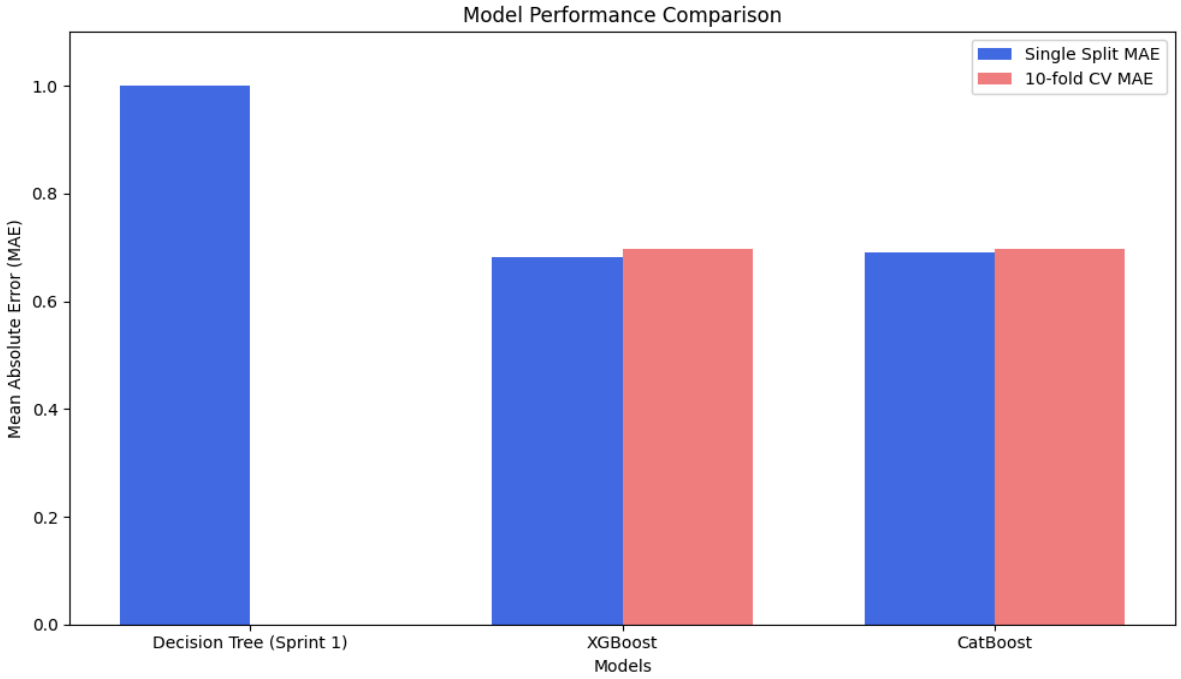
Success will be measured through the mean absolute error, or the MAE, as this metric provides an average of errors in predictions. The value for this study is currently at 1.0. If we tune the model and lower that value, we could enhance the accuracy. Another metric was determining which variables actually contribute most to mood swings. To ensure the generalizability, the data was split and tested across different subsets. The dataset I chose is a good starting point, but honestly, we would have better results and model performance if we had more expanded records. From sprint 1&2, the original decision tree model had a MAE of 1.0. XGboost had a slightly better mae value, while Catboost was just barely better on the avg 10-fold CV. The overall strength of XGBoost is that it gave us the lowest MAE, which is a significant improvement from sprint 1. However, the manual feature encoding was annoying & messy and it's not as suited for categorical data. For Catboost, the strengths include being stable/consistent as it had a better average cross validation MAE.
XGBoost produced the better MAE of 0.6821, with a tenfold cross-validated MAE of 0.6976, slightly outperforming CatBoost, which had a tenfold MAE of 0.6974
For error analysis, it was revealed that the models struggled with cases that had extreme values like days indoors or even conflicting values with variables like stress or coping struggles.

CV STD print

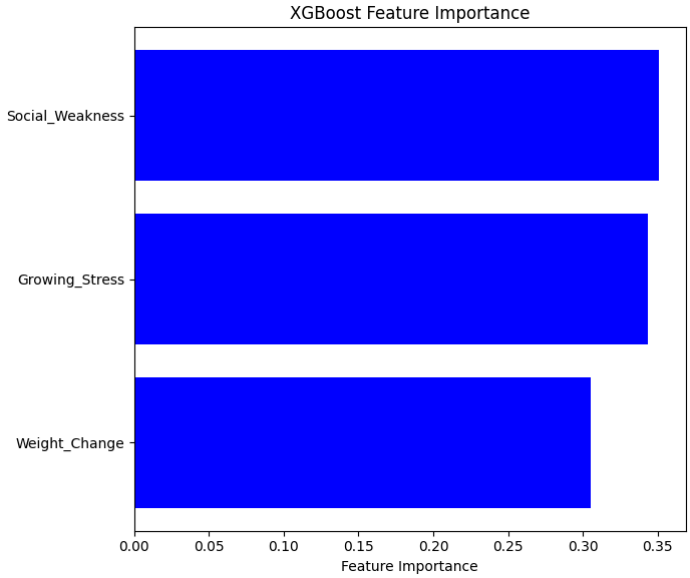
```
print("XGBoost CV STD:", np.std(xgb_cv_scores))
print("CatBoost CV STD:", np.std(cat_cv_scores))
```

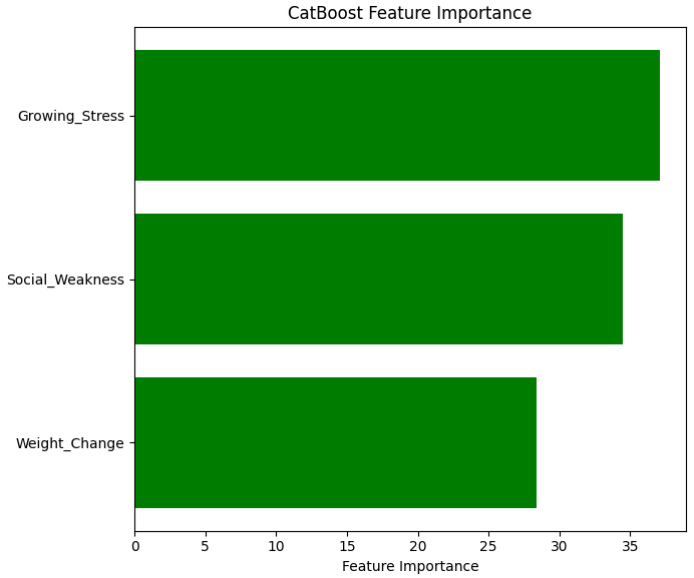
Kareem Hasan
INST414

XGBoost CV STD: 0.042CatBoost CV STD: 0.043



Feature importance plots for xgboost and catboost models below:





Kareem Hasan
INST414

Implementation Strategy and Business Value

The implementation strategy can be for a NGO to embed our optimized model into their digital health platforms or community-survey screening tools. As soon as the algorithm flags someone with high predicted mood swings, a counselor or outreach person is prompted to follow up which would enable earlier check-ins, or remote counseling sessions/workshops. The quantitative impact is that early identification can reduce the economic costs associated with untreated mental health conditions, like unhealthy coping mechanisms or antianxiety/antidepressant medication. The overall business value is the ability to transition from reactive to proactive mental health care, directly improving lives while enhancing organizational effectiveness. Technical challenges would be things like ensuring the handling of sensitive health data is encrypted and secured.

Ethical Considerations and Limitations

Some ethical considerations are Privacy & consent as collecting behavioral and self-report data requires anonymization, clear informed consent, and security to protect individuals' identities. Fairness and bias in the model is another concern since the RHMCD-20 dataset skews toward younger adults (16–30) and may under-represent other groups. Gender imbalances could also skew predictions the same way. Some limitations of our data is that it's based all on self-report like "coping struggles" as a proxy and not clinical diagnoses, so it cannot replace professional assessment. Finally, because the data comes from Bangladesh during COVID, we can't generalize in other contexts without retraining on new, representative datasets. We can mitigate this by expanding the training pool if possible and regularly evaluating the performance of the model. and making sure

References/Sources/Bibliography

Kareem Hasan
INST414

Amin, N., Salehin, I., Baten, M. A., & Noman, R. A. (2024, April 5). RHMCD-20 dataset: Identify Rapid Human Mental Health Depression during quarantine life using machine learning. Data in brief. https://pmc.ncbi.nlm.nih.gov/articles/PMC11016953/

Pelt, D. H. M., Habets, P. C., Vinkers, C. H., Ligthart, L., van Beijsterveldt, C. E. M., Pool, R., & Bartels, M. (2024, August 14). Building machine learning prediction models for well-being using predictors from the exposome and genome in a population cohort. Nature News. https://www.nature.com/articles/s44220-024-00294-2

Salehin, Imrus; Amin, Nazrul (2024), "The RHMCD-20 datasets for Depression and Mental Health Data Analysis with Machine Learning ", Mendeley Data, V2, doi: 10.17632/pxjmjyfdh2.2

Susanty, S., Sufriyana, H., Su, E. C.-Y., & Chuang, Y.-H. (2023, January 25). Questionnaire-free machine-learning method to predict depressive symptoms among community-dwelling older adults. PloS one. https://pmc.ncbi.nlm.nih.gov/articles/PMC9876369/

Vindegaard, N., & Benros, M. E. (2020, October). Covid-19 pandemic and Mental Health Consequences: Systematic review of the current evidence. Brain, behavior, and immunity. https://pubmed.ncbi.nlm.nih.gov/32485289/