May Berger & Sebastian Berlingieri

Final Project Report

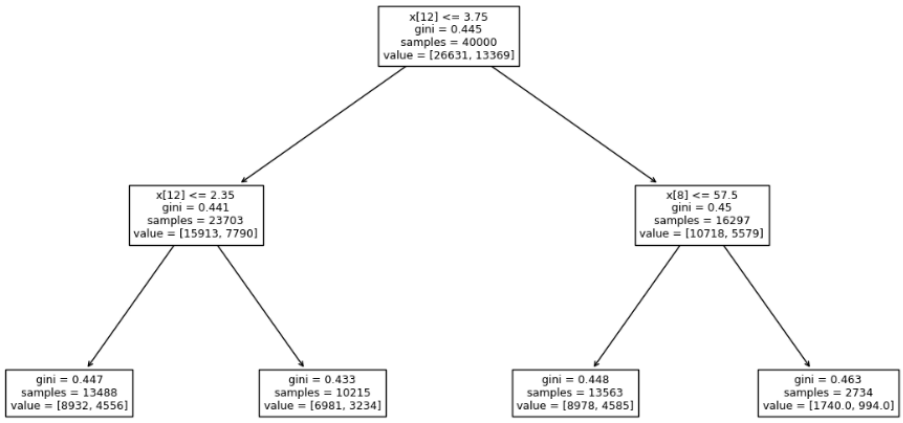
Introduction

Social media usage has increased drastically in the past few decades as technology becomes more available and new platforms are created. Lots of research has been done that predicts that social media has a negative effect on people's mental health. The stress that social media has caused on certain people could cause them to take up negative or detrimental habits such as smoking and drinking more. I predict that as social media usage increases so does the severity of users' mental health which can cause them to increase their drinking and smoking habits. Using data from a survey asking a variety of adults to answer numerous questions about themselves, their mental health, and their habits I use a decision tree (DecisionTreeClassifier) to see how well we can predict the severity of their mental health using data about their behavior. The variables we will use to predict their stress levels are social media usage, sleep hours, work hours, physical activity hours, diet quality, smoking habits, and alcohol consumption. With the wide variety of data, we were unable to get accurate results that could truly show that all of these factors did impact the severity of their mental health.

Methods

Here, the Decision Tree model was used for classifying the severity of mental health conditions by considering a dataset including predictors such as Diet_Quality, Smoking_Habit, Alcohol_Consumption, Stress_Level, Gender, etc. Ordinal versus One-Hot Encoding. The dataset was cleaned and preprocessed: ordinal features were encoded using OrdinalEncoder, and categorical features were encoded using OneHotEncoder. The target variable, severity, was encoded in binary classes for simplicity. The dataset was divided into the training and testing sets

on the 80%-to-20% ratio. In conclusion, to tune the Decision Tree model, Hyperparameters (max_depth, min_samples_split) are tuned using the GridSearchCV utility. Cross-validation accuracy was used to choose the best model. The decision rule process was visualized of the final Decision Tree model. The tree was trained on the full 46,000 samples, while the forest model was trained on the parent tree, which was calibrated on the root node as defined above. This split resulted in two child nodes: the left child node split on x[12] <= 2.35 with a Gini impurity of 0.441, covering 29,703 samples, and the right child node split on x[8] <= -5.75 with a Gini impurity of 0.448, covering 16,297 samples. The subsequent splits of the left child node nodes left_child and right_child have Gini impurities of 0.447 and 0.433, covering 13,848 and 15,855 samples, respectively. The right child node/w's splits gave nodes with Gini impurities of 0. 483 and 0.426, involving 6,121 and 10,176 samples, respectively. The splits represent where the model makes choices from features over a certain threshold that contribute to a classification.

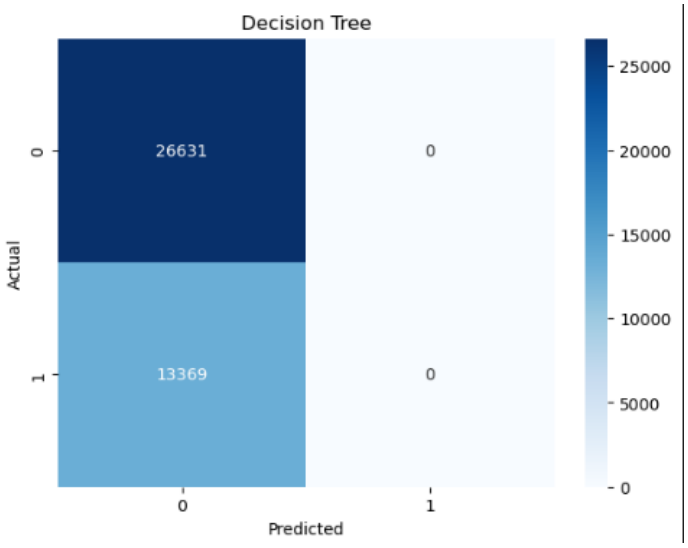


Results

The confusion matrix of the decision tree model, provided a rich understanding of its performance in predicting the severity of mental health. Figure 1 of the matrix indicates:

1. True Negatives (TN): 26,631 products where the true class was 0 and predicted as 0.
2. False Positives (FP): 0 cases where it was a member of the class but was predicted as0.
3. False Negatives (FN): 13,369 with true class as 1 while predicted as 0.
4. True Positives: 0 times where the actual class of the instance was 1 and the predicted was also 1.

The high number of true negatives reveals the good capture of the implicates of low-grade cases. But, since the number of true positives was zero and the number of false negatives was large, that means the inability to recognize high-severity patients. The decision tree is visualized using the plot_tree function and shows you how the model decides to make a prediction, for example, the first split is on feature x[12] <= 3.75) Later splits then further describe the class based on other properties, accommodating their importance. In general, the Decision Tree model appears to be effective in categorising moderate severity mental health, but has the potential to improve in the prediction of high severity. The confusion matrix and decision tree graph prove to be useful in understanding the weaknesses and strengths of the model.



Discussion

We ran into issues with missing data in the severity section which is the exact variable we were looking to predict. We overcame this issue by marking missing severity scores as low severity however this could create bias and we are unsure what exactly the true severity of these users' mental health was. There is definitely room for improvement and a data set with no missing variables could be very helpful. Social media's impact on stress levels and mental health is still a new topic where little research has been done. This project was very important and if it were to continue it could be beneficial to survey college students to see how it has impacted them. With the new survey we would emphasize the anonymity of the survey so that hopefully they will answer more truthfully and fill out every question. It may have also been helpful to delete responses from any users who did not fill out every single question. We did expect more clear results with better accuracy but once again there were issues with the data set. It may also be interesting to see if there were better results when using random forests (RandomForestClassifier). There is still lots of work and research to be done when it comes to predicting how social media affects mental health. Another angle that could be looked at in the future is how social media affects the mental health of young children or teenagers.

Works Cited

Pandey, B. (2024). Mental Health and Lifestyle Dataset for Sentiment Analysis [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14838680