Yuta Hirahata

2600220326-9

Data Science

Final Assignment: Predictive Modeling with Domain-Specific Data

Impact of Social Media Usage on Emotional States

I. INTRODUCTION

Social media is a digital technology that allows the sharing of ideas and information, including texts, images, videos, and so on. Since it was invented in 1990s, it has changed people's way of communication entirely. Now, people no longer need to take hard time to find their friends who share the same characteristics. On the other hand, this convenience has made people be addicted to social media excessively. For example, people started to regard the number of followers as an important status of their popularity or even the value of themselves. Therefore, losing their followers becomes extreme stress to them. Therefore, the academic departments should establish educational system that stops people from social media addiction. Moreover, people who are already addicted to it must learn a proper way of using those online tools. Nowadays, there are multiple social media applications, such as Instagram, X, Linked In, and so on. People use different social media platforms depending on their hobbies, personalities, and purposes. Also, the amount of time people spend on social media per day differs among their genders and ages. In this project, we investigate how the use of social media influence people's emotional states based on the following factors: users' age, gender, most used social media platform, as well as the number of posts, likes, comments, and messages per day.

II. METHODOLOGY

2.1 Dataset Explanation

Our dataset is carefully selected from Kaggle (a data science community where numerous datasets, codes, and insights can be achieved). Social Media Usage and Emotional Well-Being dataset was published by Emirhan BULUT, and it had been monthly updated until May 2024. The data was collected using a hypothetical survey-based methodology, which was intended to obtain the patterns of social media use on users' emotional states. During the survey procedure, participants were asked about their daily social media usage, then they were also asked to report their emotional state at the end of the day. Those responses are precisely collected and organized into an Excel file, which includes the following columns.

- User ID: assigning a different index to each respondent.

- Age: intended to group the participants' generation afterwards.

- Gender: each participant needs to select either male, female, or non-binary

- Platform: the most used platform among Instagram, X, Facebook, LinkedIn, Snapchat, Whatsapp, and Telegram.

- Daily Usage Time: the amount of time they spend on those social media platforms per day (in minutes).

- Posts Per Day: how many times they post on social media per day.

- Likes Received Per Day: how many likes they achieve in a day.

- Comments Received Per Day: how many comments they get per day.

- Messages Sent Per Day: how many messages they see in a day.

- Dominat Emotion: their emotional state at the end of the day.

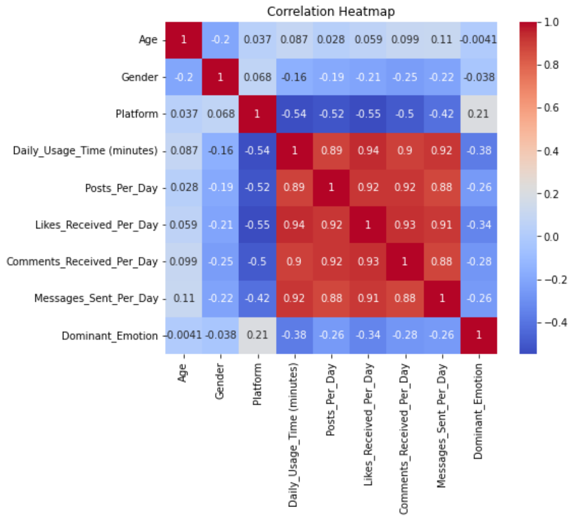
This dataset is authenticated with MIT license, thus it is used ethically and responsively in accordance.

2.2 Experimental Design

To make sure the data usability, we start the experiment with visualizing data structures, characteristics, and correlations. Then, we organize the data by deleting unnecessary columns as well as quantifying string variables. After these steps, we effectively implemented data classification using six different models.

2.2.1 Data Visualization and Organization

After reading data from the excel file, overall shape of the data is first defined. It initially contained 2000 rows and 10 columns. Next, rows and columns with no values are deleted from the table, which results in leaving 1000 rows and 10 columns. In addition, our project examines the relationships between people's daily activity on social media and their mental states, which do not require the specification of respondents. Therefore, user-id column is removed from the table. In consequence, our data classification is done with the rest of 1000 rows and 9 columns of datapoints. Furthermore, in order to make the data all quantifiable, conversion of string variables to float variables is processed. To be specific, there are initially three genders: male, female, and non-binary. We assigned an index to each gender, in which male equals to 0, female equals to 1, and non-binary equals to 2. Similarly, each of the seven social media platforms is also quantified, where Instagram equals 0, Twitter (X) equals to 1, Facebook equals to 2, LinkedIn equals to 3, Snapchat equals to 4, WhatsApp equals to 5, and Telegram equals to 6. Lastly and most importantly, people's emotional states are also converted into numbers. This includes happiness equals to 0, neutral equals to 1, boredom equals to 2, anxiety equals to 3, anger equals to 4, and sadness equals to 5. After all these quantifications, the correlations among those columns are defined (figure 1). As seen in the heatmap, there are strong correlations between some of the columns, such as the number of posts and the number of likes. However, since emotional states are not scaled from 1 to 5 (instead there are just assigned index), the direct correlation cannot be defined from this heatmap. Therefore, the next step is to classify those emotional states based on social media use activities and investigate if they are predictable with those factors.



(Figure 1)

2.2.2 Data Classification

Before any classification models are applied, the dataset is split into 80% train data and 20% test data with a random state of 269. Then, six different classification models are implemented, including four naïve bayes models, random forest, and one neural network model.

- Gaussian Naïve Bayes: Assumes that features follow a normal distribution, making it effective for continuous data.

- Multinomial Naïve Bayes: Ideal for count-based data, often used in text classification tasks.

- Bernoulli Naïve Bayes: Works with binary data, focusing on the presence or absence of features.

- Categorical Naïve Bayes: Handles categorical data by calculating probabilities for discrete feature values.

- Random Forest: Combines multiple decision trees to improve accuracy and reduce overfitting.

- Feed-Forward Neural Network: Uses a layered structure to learn patterns and relationships in complex, nonlinear data.

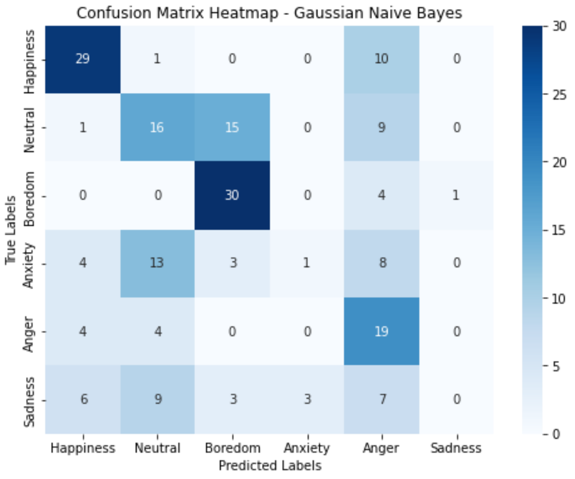
For random forest model, five different number of decision trees are applied: 5, 25, 50, 100 and 200. After the implementation, those models are precisely compared with the accuracy they achieved for predicting people's emotional states. In addition, a confusion matrix is visualized for each model's results. With confusion matrix, it is easier to distinguish between values of correctly predicted ones and values of misclassifications.

III. RESULTS

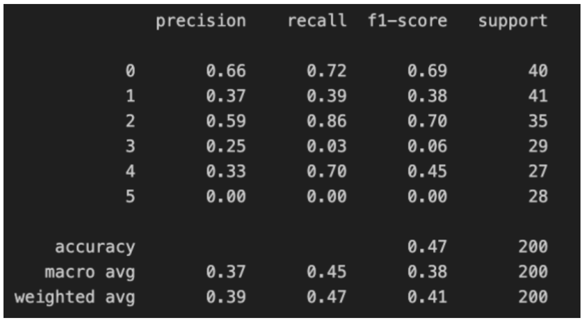
3.1 Numerical Results of Each Model

3.1.1 Gaussian Naïve Bayes

The confusion matrix indicates that happiness and boredom are predicted with high accuracy, yet other emotional states are classified incorrectly (figure 2). The overall accuracy score is 0.47 (figure 3). In other words, Gaussian Naïve Bayes model achieved 47% accuracy score for this prediction.



(Figure 2)

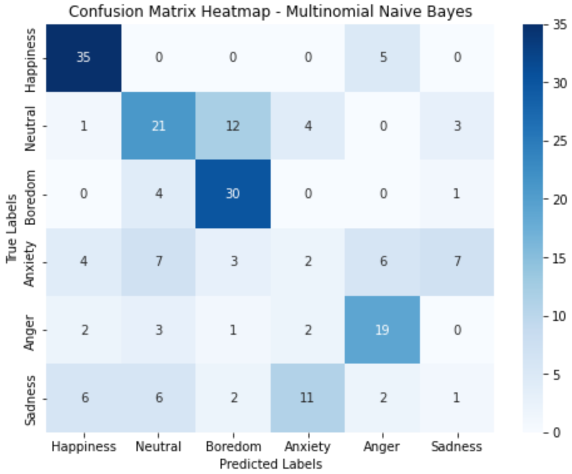


```
              precision    recall  f1-score   support

           0       0.66      0.72      0.69        40
           1       0.37      0.39      0.38        41
           2       0.59      0.86      0.70        35
           3       0.25      0.03      0.06        29
           4       0.33      0.70      0.45        27
           5       0.00      0.00      0.00        28

    accuracy                           0.47       200
   macro avg       0.37      0.45      0.38       200
weighted avg       0.39      0.47      0.41       200
```

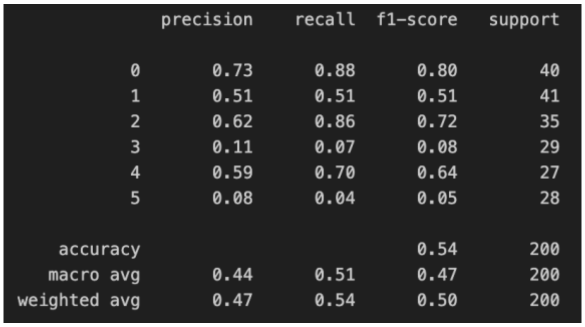
(Figure 3)

3.1.2 Multinomial Naïve Bayes

The results for this model is a little similar to the results from Gaussian Naïve Bayes. Happiness and Boredom are appropriately predicted, and neutral and anger obtain relatively high accuracy compared to other emotional (figure 4). The overall accuracy for Multinomial Naïve Bayes model is 0.54, meaning the 54% accurate prediction (figure 5).



(Figure 4)

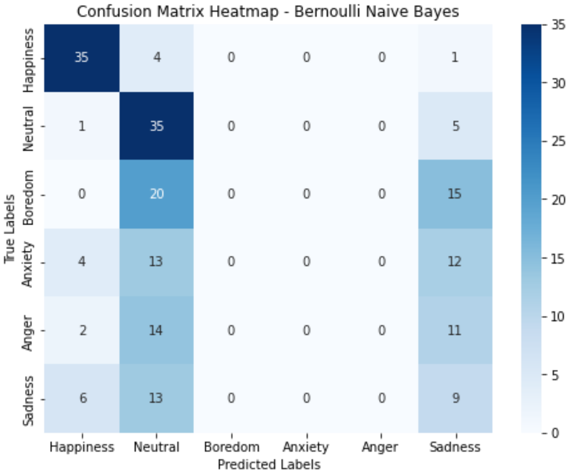


```
              precision    recall  f1-score   support

           0       0.73      0.88      0.80        40
           1       0.51      0.51      0.51        41
           2       0.62      0.86      0.72        35
           3       0.11      0.07      0.08        29
           4       0.59      0.70      0.64        27
           5       0.08      0.04      0.05        28

    accuracy                           0.54       200
   macro avg       0.44      0.51      0.47       200
weighted avg       0.47      0.54      0.50       200
```

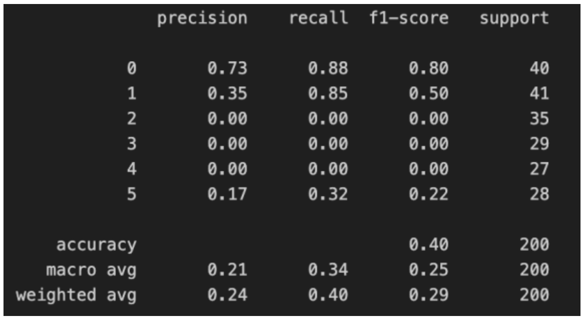
(Figure 5)

3.1.3 Bernoulli Naïve Bayes

This model has the lowest prediction accuracy among the six models used in this project. Happiness and neutral are predicted correctly, but other emotional states are terribly predicted. Boredom, anxiety, and anger are treated as if there are no such emotional states. Instead, most of them are predicted as neutral (figure 6). The overall accuracy score is 0.40, hence it correctly predicts only 40% of the data (figure 7).



(Figure 6)

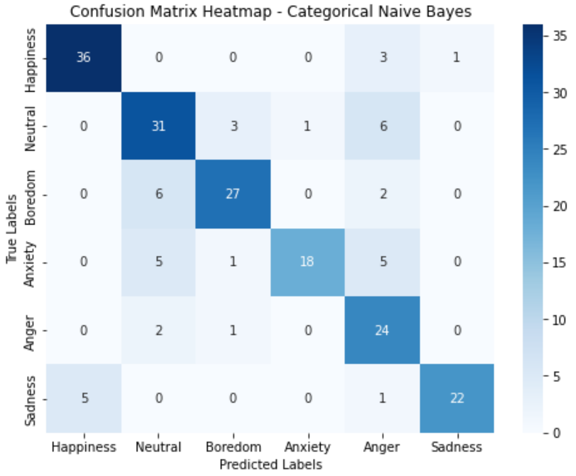


```
              precision    recall  f1-score   support

           0       0.73      0.88      0.80        40
           1       0.35      0.85      0.50        41
           2       0.00      0.00      0.00        35
           3       0.00      0.00      0.00        29
           4       0.00      0.00      0.00        27
           5       0.17      0.32      0.22        28

    accuracy                           0.40       200
   macro avg       0.21      0.34      0.25       200
weighted avg       0.24      0.40      0.29       200
```

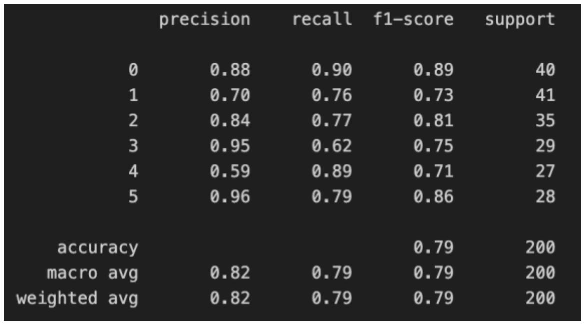
(Figure 7)

3.1.4 Categorical Naïve Bayes

Among the four Naïve Bayes models, this model results in the best accuracy. As seen in the confusion matrix (figure 8), diagonal cells have the darkest blue colors compared to other cells. This indicates that the model has predicted each emotional state appropriately. However, there are still some misclassifications seen in the matrix. The numerical accuracy score for Categorical Naïve Bayes model is 0.79 (figure 9), which is much higher score than other Naïve Bayes models.



(Figure 8)

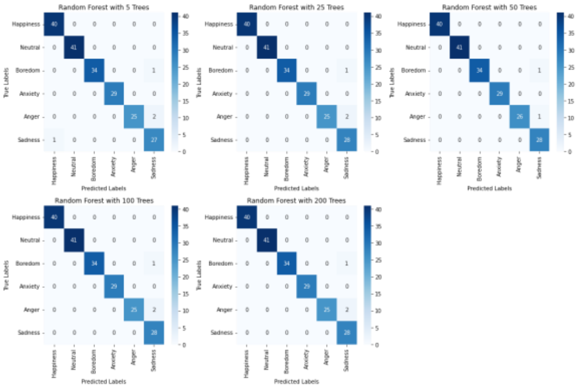


```
              precision    recall  f1-score   support

           0       0.88      0.90      0.89        40
           1       0.70      0.76      0.73        41
           2       0.84      0.77      0.81        35
           3       0.95      0.62      0.75        29
           4       0.59      0.89      0.71        27
           5       0.96      0.79      0.86        28

    accuracy                           0.79       200
   macro avg       0.82      0.79      0.79       200
weighted avg       0.82      0.79      0.79       200
```

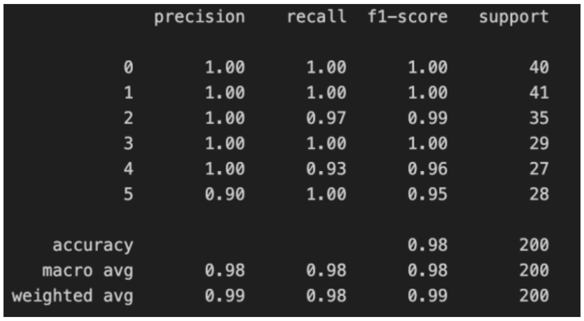
(Figure 9)

3.1.5 Random Forest

To investigate what the appropriate range of patterns the random forest model should achieve, we prepared five different number of trees. The confusion matrices below (figure 10) represent the performance of random forest with each number of trees. All of them performed very well, so the diagonal cells have the darkest blue colors. Only the one with 50 trees has predicted the anger emotion slightly better than others. All the other decision tree values obtained exactly the same results. Therefore, the quantified results shows that random forest model with 5, 25, 100, and 200 trees perform the prediction with 98% accuracy (figure 11), while that with 50 trees performs the prediction with 99% (figure 12).

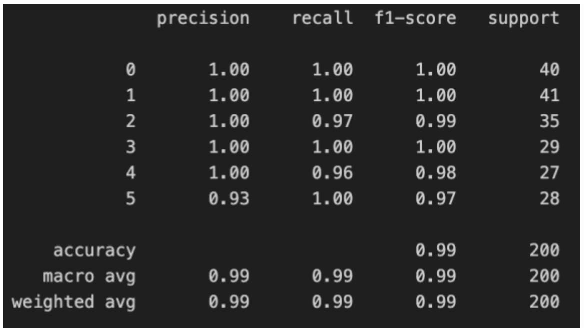


(Figure 10)



```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        40
           1       1.00      1.00      1.00        41
           2       1.00      0.97      0.99        35
           3       1.00      1.00      1.00        29
           4       1.00      0.93      0.96        27
           5       0.90      1.00      0.95        28

    accuracy                           0.98       200
   macro avg       0.98      0.98      0.98       200
weighted avg       0.99      0.98      0.99       200
```

(Figure 11)

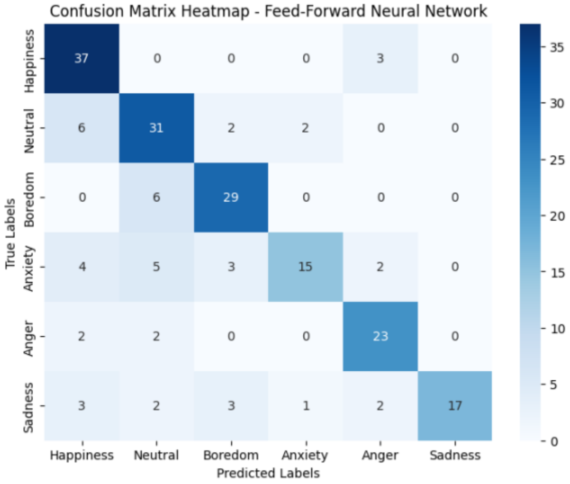


```
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        40
           1       1.00      1.00      1.00        41
           2       1.00      0.97      0.99        35
           3       1.00      1.00      1.00        29
           4       1.00      0.96      0.98        27
           5       0.93      1.00      0.97        28

    accuracy                           0.99       200
   macro avg       0.99      0.99      0.99       200
weighted avg       0.99      0.99      0.99       200
```

(Figure 12)

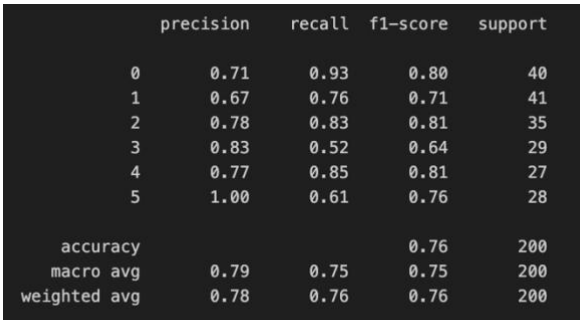
3.1.6 Feed-Forward Neural Network

This model also achieved relatively high accuracy score. As seen in the confusion matrix (figure 13), diagonal cells are properly shown with darker blue colors than other cells, indicating appropriate predictions. Nevertheless, it takes hard time to predict anxiety and sadness emotions compared to other mental states. The overall accuracy score is 76% (figure 14).



	Happiness	Neutral	Boredom	Anxiety	Anger	Sadness
Happiness	37	0	0	0	3	0
Neutral	6	31	2	2	0	0
Boredom	0	6	29	0	0	0
Anxiety	4	5	3	15	2	0
Anger	2	2	0	0	23	0
Sadness	3	2	3	1	2	17

(Figure 13)



```
              precision    recall  f1-score   support

           0       0.71      0.93      0.80        40
           1       0.67      0.76      0.71        41
           2       0.78      0.83      0.81        35
           3       0.83      0.52      0.64        29
           4       0.77      0.85      0.81        27
           5       1.00      0.61      0.76        28

    accuracy                           0.76       200
   macro avg       0.79      0.75      0.75       200
weighted avg       0.78      0.76      0.76       200
```

(Figure 14)

3.2 Summary

Overall, the Naïve Bayes variants demonstrated a range of outcomes, with Gaussian Naïve Bayes (47%), Multinomial Naïve Bayes (54%), and Bernoulli Naïve Bayes (40%) exhibiting relatively modest accuracy. In contrast, the Categorical Naïve Bayes model achieved the highest performance within the Naïve Bayes family (79%). Random Forest outperformed all other approaches, reaching a nearly perfect prediction accuracy—98% for 5, 25, 100, and 200 trees, and 99% for 50 trees. Finally, the Feed-Forward Neural Network also demonstrated solid performance, attaining a 76% accuracy rate.

IV. DISCUSSION

Each classification model has different characteristics, so the most appropriate model choice varies among datasets. It not only relies on data types but also data size and dimension. Social Media Usage and Emotional Well-Being dataset has 1000 rows, which is typically regarded as a small dataset. However, it includes 9 columns of data features, so the correlations are so complicated. Moreover, figure 1 shows that there are some columns having strong correlations while others having weak correlations. In this part, we consider all of those factors to precisely analyze the results.

4.1 Naïve Bayes Comparison

To begin with, Categorical Naïve Bayes models performed the best compared to other three Naïve Bayes models.

Gaussian Naïve Bayes model assumes that features follow a normal distribution, so it is suitable for continuous data. In contrast, our dataset is a discrete data, and some columns' values are manually converted from string to float. This incompatibility has resulted in low accuracy score for such model. Multinomial Naïve Bayes model is often used for count-based data, such as text classification tasks. Unlike Gaussian Naïve Bayes model, it can manage discrete data, but our data is not count-based. Consequently, it achieved low accuracy value yet performed slightly better than Gaussian Naïve Bayes. Bernoulli Naïve Bayes is highly compatible with binary data while our dataset has eight input features and six classification outputs. This critical difference between ideal data and our data led to such model having the lowest performance among all models.

On the other hand, Categorical Naïve Bayes model is literally capable of categorical data by calculating probabilities for discrete feature values. Therefore, this model performing the best means our dataset has more categorical data than numerical data. Nonetheless, the accuracy score of this model is still 79%. This is caused by some columns containing numerical values. Also, Categorical Naïve Bayes performs the best when each feature is completely independent. However, figure 1 shows some columns having strong relationships, meaning one is dependent of others. Accordingly, a model which can handle data containing both categorical and numerical values; furthermore, that model must be irrelevant to feature correlations.

4.2 Random Forest

Random forest is an integration of multiple decision trees, which enhances the accuracy as well as reduces overfitting. It performs the best when dataset is well-structured, and it can handle both categorical and numerical data. Also, the strength of correlation among features does not affect its results much. Therefore, random forest is highly suitable for our dataset.

Also, it is important to set an appropriate decision tree values, and we tested out five different values: 5, 25, 50, 100 and 200. Among those five values, one with 50 trees performed slightly better than others. When a small number of trees is applied, it might result in reducing the effectiveness of the random forest model because of the insufficient diversity in the ensemble model. In other words, predictions are more likely to be influenced by the bias or noise of individual decision trees. This unstable condition led the model with 5 and 25 trees to obtain lower accuracy score than that with 50 trees. On the other hand, when the number of trees is too large, it improves the model's robustness and reduces variance, but at the same time, it may lead to redundant computations and overfitting. Especially, this problem arises when the dataset is not large enough to justify such complexity. Since our dataset is not so large, occurrence of such issue is highly possible. This must have caused the model with 100 and 200 trees to perform less than that with 50 trees. According to these insights, it can be concluded that random forest model with 50 trees has an ideal trade-off between bias and variance. This allowed the model to capture the underlying patterns of the dataset without overfitting or adding unnecessary computational cost.

4.3 Feed-Forward Neural Network Model

This model is capable of learning complex patterns and relationships in data using multiple layers of neurons; thus, it is compatible with our dataset having high-dimensional structure. However, since this model needs to process a series of deep learning tasks, it requires a large amount of data to generalize well. In such case, our dataset is lack of its size to be compatible with this model; hence, this model obtained 76%, which is lower than the results achieved from random forest and Categorical Naïve Bayes models.

4.5 Summary

Each classification model's characteristics and our dataset's structures have huge impacts on the project results. Although Categorical Naïve Bayes model has shown the highest quality of prediction accuracy, the presence of numerical data as well as strong correlations among some features restricted its performance. Conversely, Random Forest model could handle both numerical and categorical data, and it is resistant to the variance of feature correlations, it predicted people's emotional states the most accurately among the six models used in this project. In particular, Random Forest model with 50 trees has achieved 99% prediction accuracy, which outperformed the others. Lastly, while Feed-Forward Neural Network can learn complex patterns in data, it requires large number of datapoints, which our dataset does not have enough this time. This underscores the necessity of aligning model choice with dataset characteristics. This discussion reinforces the need to carefully consider data structure, size, and feature interactions when selecting machine learning models.

V. CONCLUSION

This project proves that the use of social media strongly affects people's emotional states. With age, gender, platform, and daily usages (the amount of time, number of likes, posts, comments, and messages) only can predict people's mental states with 99% accuracy, which indicates extremely high addiction.

It is true that social media has established a new opportunity for people to connect with others and enrich their daily lives. However, if the use of such tools makes people to get stressed or depressed, they become totally inconvenient.

In conclusion, it is essential for educational institutions to teach students and teachers so that they can reduce their dependency on social media. IN addition, for adults who are already addicted to social media, it is substantial to use social media to the extent that they can control their emotions, such as by consciously reducing the amount of time they spend on it.

Reference

[1] BULUT, E. (2024, May 19). Social media usage and emotional well-being. Kaggle. https://www.kaggle.com/datasets/emirhanai/social-media-usage-and-emotional-well-being?select=train.csv