Greenspace Team 3 - Therapeutic Alliance
Week 4 Team Report

Kohsin Chen, Zerui Zhang, Zheng Zhang, Bingshen Yang

- Team progress compared to the project plan and millstones

In order to find more predictors for therapeutic alliance, our team has decided to undertake the following division of tasks to further explore and optimize the model:
Kohsin is tasked with determining the main diagnosis based on the types of assessments assigned to each patient, calculating the assessment completed ratio based on the dataset.
Zerui is responsible for identifying patient grouping information in the tags. This will help consider the characteristics of different patient groups in the model.
Bingshen and Zheng are in charge of filtering through the questions and responses to identify characteristics of patients that may impact their therapeutic alliance scores.

In the end, the independent variables identified were: PATIENT_TYPE, INOFFICE, SUICIDALITY_FLAG, AVERAGE_INITIAL_ASSESSMENT_SCORE, COMPLETED_RATIO, DIAGNOSIS, AGE, GENDER, and OCCUPATION. The dependent variable is THERAPEUTIC_ALLIANCE_SCORE.

Result: The random forest model, where NaN values were imputed with the most frequent value, showed a slight improvement over other methods but still performed weakly, with a Mean Squared Error (MSE) of 192.77 and an adjusted R-squared of 0.317. In contrast, the random forest model with dropped NaN values yielded an MSE of 220.18 and an adjusted R-squared of 0.193. We also experimented with using Lasso for feature selection prior to applying the random forest model, but the outcome remained unsatisfactory, resulting in an MSE of 222.383 and an adjusted R-squared of 0.192. We hypothesize that the poor performance might be attributed to the limited information in the dataset.

According to the project plan, Kohsin, Zerui, Bingshen, and Zheng are involved in the model setup and fine-tuning, which is scheduled to be completed by May 26, 2024. This part of the work is progressing as planned. However, we have determined that the current data is unable to predict therapeutic alliance outcomes, as all models are performing poorly. Next step we will focus on the therapist match.

- The individual contributions on what they have done in previous week

Group Member	Contribution	Challenges
Kohsin	1. Extracted the data to determine the assessment completion ratio for each patient, defined as the number of completed assessments divided by the total number of assessments assigned to a patient. 2. Identified the primary diagnosis based on the types of assessments assigned to each patient. Since identifying the primary diagnosis might be challenging, I followed Greenspace's suggestion to classify patients based on the most frequently given assessments to each patient. 3. Incorporated all the identified variables and built Random Forest and LASSO models to investigate if there is a predictor for therapeutic alliance (TA).	Identifying the primary diagnosis is arbitrary. Using this method to identify the primary diagnosis might be problematic, but due to the lack of such information in the dataset, it is a necessary compromise. Some assessments are too general to classify the patient. For example, some questionnaires measure daily function, which does not allow for clear classification. As a result, many patients will still have missing values for their main diagnosis.
Zerui	1. Identified over 400 mental health conditions tags and 400 treatment methods tags. Extracted key information from complex tags to create new data columns. 2. Used keyword matching to sort tags into main diagnosis and treatment categories.	This week's main challenge was dealing with the large volume of tags, most of which were not directly useful. Sorting out the relevant tags required a lot of manual work. Additionally, preparing the data for modeling was complicated by missing data, duplicates, and the need for

	Merged these categories with the patient table. 3. Developed a Random Forest Regressor model by combining these tags with demographic and assessment data, evaluating it based on R-squared, MSE, and feature importance.	extensive merging of datasets.
Zheng	1. Identify the most frequent condition question from big clinics. Classified and characterized patients using selected questions and their response data. 2. Built the regression model and random forest model predict therapeutic alliance scores using attributes from responses, and final selected attributes from all tables. Utilized LASSO and forward/backward selection to adjust the variables selection.	Lots of NAN value, since the questions from big clinics won't be asked by other clinics. Deciding between different models and determining their configurations can be challenging due to the trade-offs between bias, variance
Bingshen	1. Further extracted key patient characteristics from the dataset based on questions and responses (two tables in the databse), besides the patients' age, gender, and occupation from last week, added Mother Tongue, Ethnicity, Suicide Attempt, Violent Behaviour, Deliberate Self-Harm, Homicidal Threats, Experiencing Violence. 2. Again, constructed models to forecast therapeutic alliance scores from assessment data.This time, I tried linear regression, random forest, Lasso, Ridge and Polynomial	All models were not good. The features are not informative. Imputing NaN value is another challenge.

	fitting. 3. Since the previous features are not informative and full of noise, at the end we tried to only use 'COMPLETED_RATIO', 'DIAGNOSIS', 'AVERAGE_INITIAL_SCORE', 'SUICIDALITY_FLAG', 'INOFFICE', 'Age', 'Gender', 'Occupation' as features and fine tune the models.	

- Team communication and collaboration

1. We schedule regular team meetings every Wednesday and Saturday, with occasional additional meetings on Thursday, to discuss progress, challenges, and next steps.
2. We share regular updates on everyone's progress during these meetings.
3. We discuss insights from our meetings with Greenspace and split our tasks accordingly.

- Clear work plan with tasks assigned to each person for the next week

Group Member	Next Week Tasks Assigned
Kohsin	Using SQL, extract all therapists from the largest clinic, link them to their assigned patients and their assessment responses, remove responses with a score of NA, and organize the data for other team members to analyze. Built unsupervised models for identifying undiscovered features.
Zerui	Calculate the average improvement score for patients from their first assessment to the last one.
Zheng	Identify the number of "success" patients for each therapist from the largest clinic.

Bingshen	Identify which patient type each therapist excels in treating using the dataset provided by Kohsin.Built unsupervised models suggested by instructors.