Generated by AI. Make sure to check for accuracy.

Meeting notes:

* **Introduction and Data Loading:** Bence introduced the meeting and shared a notebook for exploratory data analysis and data preprocessing. He explained the changes made to the data loading process to allow running the notebook locally or on Kaggle.
  + **Meeting Introduction:** Bence started the meeting by acknowledging Jennifer for starting the recording and mentioned that he would share his screen to present some charts and updates made since the last call. He noted that the notebook could be used on Kaggle or locally for further updates and improvements.
  + **Data Loading Changes:** Bence explained that the introduction section of the notebook remained unchanged, but he made changes to the data loading process for operational reasons. He highlighted the importance of setting a variable to either 'local' or 'Kaggle' to ensure the notebook runs correctly in different environments. This change was made to address the issue of saving notebooks on Kaggle, which sometimes causes recalculations.
* **Exploratory Data Analysis:** Bence presented the exploratory data analysis section, highlighting the importance of understanding the data and identifying which columns to keep. He discussed the differences between ADHD diagnosed and non-diagnosed individuals, as well as the gender distribution in the dataset.
  + **EDA Overview:** Bence introduced the exploratory data analysis (EDA) section, emphasizing the importance of understanding the data and identifying which columns to keep. He mentioned that the EDA section includes various charts and analyses to help visualize the data.
  + **Data Import:** Bence explained the process of importing data into a single data frame, which includes categorical, quantitative, and MRI data. He noted that the data frame is large, with 5 rows and almost 20,000 columns, primarily due to the brain cell connection matrices.
  + **Target Variables:** Bence discussed the target variables, ADHD outcome and gender, and highlighted the imbalance in the dataset. He noted that most individuals in the training dataset are diagnosed with ADHD and that there are more males than females, which could impact the model's predictions.
  + **Exploratory Charts:** Bence presented various charts to visualize the data, including count plots and frequency charts for categorical variables. He used these charts to identify differences between ADHD diagnosed and non-diagnosed individuals and to understand the distribution of different features in the dataset.
* **Categorical Variables Analysis:** Bence analyzed the categorical variables, showing the differences in frequencies between ADHD diagnosed and non-diagnosed individuals. He identified variables that should be dropped due to their lack of overlap between the training and test datasets.
  + **Frequency Analysis:** Bence presented frequency charts for categorical variables, highlighting differences between ADHD diagnosed and non-diagnosed individuals. He noted significant differences in certain features, such as race and ethnicity, and emphasized the importance of keeping these variables for further analysis.
  + **Variable Overlap:** Bence compared the distribution of categorical variables between the training and test datasets. He identified variables with no overlap, such as enrollment year and study site, and recommended dropping these variables as they would not be applicable to the test dataset.
  + **Chart Improvements:** Bence mentioned that he improved the charts by including all categorical variables and using frequency charts instead of count plots. This change made it easier to compare the distribution of features between ADHD diagnosed and non-diagnosed individuals.
* **Quantitative Variables Analysis:** Bence examined the quantitative variables, focusing on the differences between ADHD diagnosed and non-diagnosed individuals. He highlighted the importance of understanding the correlations between variables and the need to reduce multicollinearity.
  + **Quantitative Variables:** Bence discussed the analysis of quantitative variables, including age and emotional problems. He highlighted the differences between ADHD diagnosed and non-diagnosed individuals, noting that diagnosed individuals tend to have more emotional problems.
  + **Correlation Analysis:** Bence presented a correlation matrix for quantitative variables, showing the relationships between different features. He identified a block of closely correlated variables related to emotional problems and suggested considering multicollinearity in the modeling process.
  + **Variable Selection:** Bence explained the process of selecting quantitative variables to keep for further analysis. He used visualizations to identify variables with significant differences between ADHD diagnosed and non-diagnosed individuals and decided to drop variables with no significant differences, such as handedness.
* **MRI Data Analysis:** Bence discussed the analysis of MRI scan data, explaining the challenges of working with a large number of columns. He presented the differences in brain connectivity between ADHD diagnosed and non-diagnosed individuals and the need for dimensionality reduction techniques.
  + **MRI Data Overview:** Bence explained the structure of the MRI scan data, which includes almost 20,000 columns representing the connections between 200 brain cells. He noted the challenges of working with such a large dataset and the need for dimensionality reduction techniques.
  + **Brain Connectivity:** Bence presented charts comparing brain connectivity between ADHD diagnosed and non-diagnosed individuals. He noted that diagnosed individuals tend to have higher connectivity intensity between brain cells, but there are some exceptions that need to be considered.
  + **Dimensionality Reduction:** Bence discussed the use of principal component analysis (PCA) to reduce the dimensionality of the MRI data. He noted that the first two principal components only explain a small percentage of the variance, suggesting the need for further research to find better methods for dimensionality reduction.
* **Data Preprocessing:** Bence explained the data preprocessing steps, including handling missing data, scaling numerical variables, and encoding categorical variables. He emphasized the importance of these steps for improving the accuracy of machine learning models.
  + **Handling Missing Data:** Bence discussed the process of handling missing data, noting that the dataset is relatively clean with only a few missing values. He explained the strategies for imputing missing values, such as using the most frequent value for categorical variables and the median for quantitative variables.
  + **Scaling Numerical Variables:** Bence explained the importance of scaling numerical variables to ensure they are on the same scale. He described the use of the standard scaler method, which standardizes variables based on their mean and standard deviation.
  + **Encoding Categorical Variables:** Bence discussed the process of encoding categorical variables using one-hot encoding. This technique creates binary variables for each category, allowing the model to capture specific features for different categories.
* **Model Training and Results:** Bence shared the results of the model training, noting improvements in accuracy after excluding certain features. He identified issues with the gender prediction model and suggested further research to improve the model's performance.
  + **Model Training:** Bence explained the process of training the model using the lightGBM algorithm. He noted that excluding certain features improved the model's accuracy from 0.84 to 0.85.
  + **Gender Prediction Issues:** Bence identified issues with the gender prediction model, which predicted all individuals as male. He suggested further research to address this issue and improve the model's performance.
  + **Future Improvements:** Bence recommended exploring different models and combining them using an ensemble approach to enhance accuracy. He emphasized the need to improve the gender prediction model and consider dimensionality reduction techniques for the MRI data.
* **Submission and Next Steps:** Bence demonstrated how to submit the model's predictions to the Kaggle competition and discussed the importance of improving the gender prediction model. He suggested trying different models and combining them to enhance accuracy.
  + **Submission Process:** Bence demonstrated the process of submitting the model's predictions to the Kaggle competition. He explained how to generate the submission file and submit it to the competition to receive a score for the test dataset.
  + **Next Steps:** Bence discussed the next steps, including improving the gender prediction model and exploring different models to enhance accuracy. He suggested trying an ensemble approach to combine different models and achieve better results.
* **Questions and Discussion:** Deirdre asked about data credibility in certain categories, and Bence explained the importance of considering the number of observations when evaluating the validity of data. He also mentioned the potential for using more sophisticated methods for handling missing data.
* **Conclusion and Recommendations:** Bence concluded the meeting by recommending that participants review the notebook and understand the steps involved in the analysis. He encouraged them to use AI assistance if needed and to focus on improving the model's accuracy.

Follow-up tasks:

* **Notebook Upload:** Upload the notebook to GitHub for easier access by the team. (Bence)
* **PCA Analysis:** Research and implement a more effective dimensionality reduction technique, potentially increasing the number of principal components to improve the model. (Bence)
* **Gender Prediction Model:** Investigate and resolve the issue with the gender prediction model, which currently predicts all individuals as male. (Bence)
* **Model Comparison:** Try different models and consider combining them using an ensemble approach to improve overall accuracy. (Bence)
* **Data Credibility:** Review the credibility of data categories with low observation counts and determine if they should be included in the model. (Deirdre)
* **CatBoost Algorithm:** Consult with Groania about the CatBoost algorithm and its handling of categorical variables and small categories. (Bence)