**Project Analysis Log**

**Objective**

The goal of this project is to predict **Average Stress and Emotional Overload** based on textual features, pregnancy-related variables, and sentiment analysis scores derived from Reddit posts.

**Initial Steps**

1. **Dataset Overview:**
   * The dataset consists of 100 rows with 27 columns, including:
     + Textual data (text column).
     + Sentiment scores (neg, neu, pos, compound).
     + Pregnancy-related variables such as Pregnancy Week (Model).
     + Stress intensity and emotional overload scores.
   * Target variable: **Average Stress and Emotional Overload**.
2. **Feature Engineering:**
   * TF-IDF vectorization of the text column to extract important words.
   * Additional features:
     + Text length (text\_length).
     + Number of question marks in the text (question\_marks).
     + Ratio of negative to positive sentiment (neg\_pos\_ratio).
3. **Modeling Approaches:**
   * Linear Regression.
   * Random Forest Regressor.
   * Recursive Feature Elimination (RFE) to select the top 10 features.

**Results and Observations**

**Model Evaluation**

1. **Linear Regression:**
   * Mean Squared Error (MSE): **6.79**.
   * R²: **-0.29** (negative R² indicates poor performance).
2. **Random Forest Regressor:**
   * Mean Squared Error (MSE): **8.54**.
   * R²: **-0.62** (indicating even worse performance compared to Linear Regression).

**Conclusion:** Both models failed to capture meaningful relationships between the features and the target variable. This suggests a need to improve feature selection, engineering, or try alternative models.

**Feature Importance (Random Forest):**

1. The top features selected by the model include words derived from TF-IDF:
   * from, since, first, on, up.
   * Importance scores indicate these features contribute the most to the model.
2. Pregnancy-related variables such as Pregnancy Week (Model) were **not selected as important** by RFE or Random Forest.

**Interpretation:**

* The lack of significance for Pregnancy Week (Model) suggests it may not have a strong linear or nonlinear relationship with the target variable. However, further investigation with engineered interactions or advanced models could provide more insight.

**Visualizations**

1. **Feature Importance Plot:**
   * Highlights the contribution of each feature, with TF-IDF words dominating.
2. **Predicted vs Actual Values (Linear Regression & Random Forest):**
   * Both plots show predictions far from the perfect prediction line, reflecting poor model performance.

**Next Steps**

**1. Try Advanced Models**

* Implement Gradient Boosting models such as:
  + XGBoost.
  + LightGBM.
* Investigate their ability to capture nonlinear relationships.

**2. Feature Engineering Improvements**

* Test interaction features:
  + Combining Pregnancy Week (Model) with sentiment scores.
  + Adding temporal features if timestamps are available.
* Perform additional text preprocessing:
  + Extract specific pregnancy-related keywords.
  + Use semantic embeddings (e.g., BERT) for richer text representation.

**3. Refine Target Variable**

* Analyze the target distribution and ensure it aligns with the prediction goals.
* Explore alternate targets such as binary stress classification (e.g., High vs Low).

**4. Reassess Pregnancy Week**

* Conduct focused analysis to determine:
  + Correlations between Pregnancy Week (Model) and the target.
  + Whether it becomes significant after adding interaction terms or nonlinear transformations.

**Challenges and Considerations**

1. Small dataset size (100 rows):
   * Limited data could restrict model performance.
   * Augmenting the dataset or using transfer learning approaches could help.
2. High-dimensional TF-IDF features:
   * Reducing dimensionality (e.g., PCA) or using domain-specific word embeddings might improve results.
3. Low R² scores:
   * Indicate the current features may not explain the target well, requiring deeper feature exploration.

**Conclusions**

* **Pregnancy Week (Model):**
  + Currently, it appears non-significant, but this does not rule out its potential if combined with other features or explored using nonlinear methods.
* **Next Actions:**
  + Focus on improving feature engineering.
  + Experiment with advanced models.
  + Document all observations to refine the process.