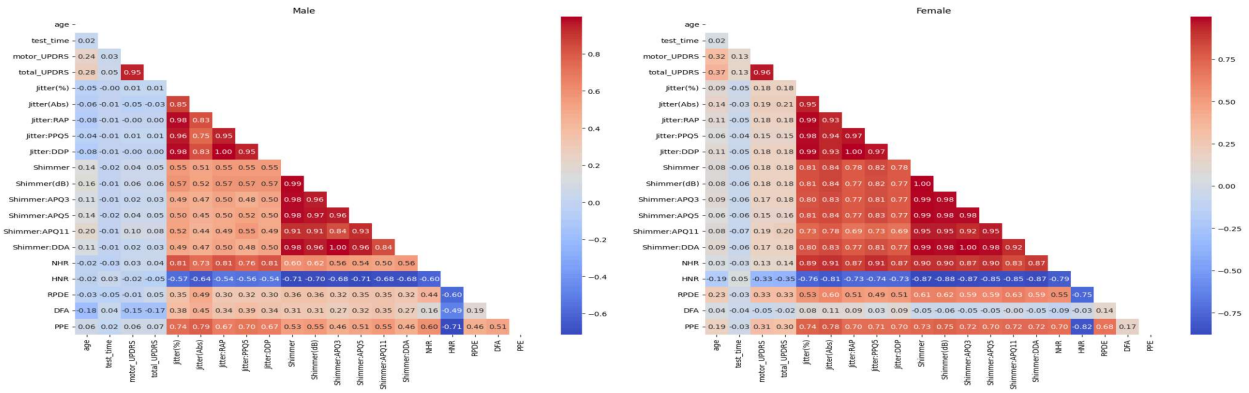
MILESTONE 2
-Harshit Soni (hs5666)

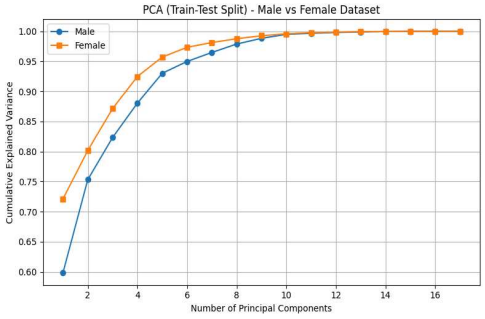
Methodology:

1. Correlation Analysis: The initial correlation showed an expected correlation between features and target variables. However, The dataset was then stratified by sex, creating male and female subgroups to evaluate whether vocal biomarkers relate differently to disease severity in each group.

1.1. Stratified Correlation: Correlation coefficients were calculated between each voice feature and both the total and motor UPDRS scores, separately for males and females. Features such as PPE, RPDE, and DFA showed strong positive correlation in both groups, while HNR had a strong negative correlation. The strength of these correlations differed by sex. For example, HNR had a correlation of -0.35 with total UPDRS in males versus only -0.05 in females. Similarly, RPDE, PPE, and Jitter(Abs) were more strongly correlated with motor and total UPDRS in males. These results highlight sex-specific differences in how vocal features associate with Parkinson's severity.



2. Data Split: To prevent data leakage, the dataset was first split in training and testing in the ratio 80:20 and then the subsequent steps for PCA were performed.

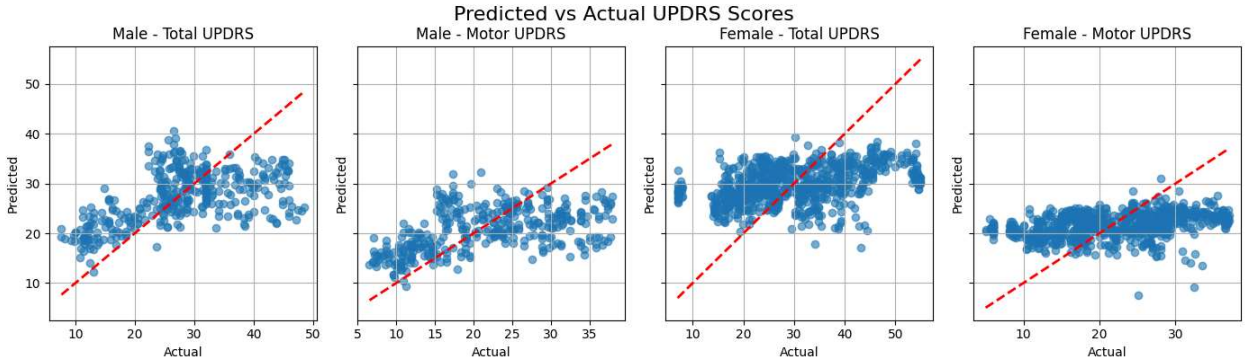


3. PCA: PCA was performed separately for male and female training datasets after standardization. This helped reduce dimensionality and multicollinearity while preserving interpretability. For males, the first 8 components captured approximately 97.5% of the variance, while for females it required 9 components to reach a similar threshold. These selected components were used in subsequent modeling for both groups to ensure comparability.

4. Linear Regression on PCA Components: Separate linear regression models were trained using the top 8 PCA components for males and 9 PCA components for females. The models predicted the clinician-rated total UPDRS (Unified Parkinson's Disease Rating Scale) score and motor UPDRS.

	Male - Total UPDRS (PCA - 97.5% - 8)	Male - Motor UPDRS (PCA - 97.5% - 8)	Linear Regression on Females (PCA - 97.5% - 9)	Female - Motor UPDRS (PCA - 97.5% - 9)
Weights	[0.63518774, 0.79372861, 3.1855923, 2.14518232, 1.73673037,	[0.54162337, 0.50679508, -2.5174969 , 1.71525485, -2.07582477,	[0.0660324 , 0.45968265, 1.5908951 , -2.83276614, -0.37906682, 0.45554236, 2.77649124,	[0.02003455, 0.34420759, 1.18597322, -1.56423352, -0.44935402, 0.3882869,

	0.9727425, -3.4350442, -3.28780726]	-1.06174852, 3.0186904 , -1.013344]	-0.28413312, -0.73422998]	2.70000672, -0.4295105, 0.40861964]
Intercept	27.441903081044874	20.808399866041526	29.649652557704304	21.449005708047412
Evaluation	RMSE: 8.216548978033401 R^2: 0.28405843221504345	RMSE: 7.321501720545059 R^2: 0.24401379795727218	RMSE: 10.668212646670451 R^2: 0.1439822929731328	RMSE: 7.560140937610359 R^2: 0.11226835293219406



Results:

1. Correlation analysis revealed sex-specific patterns in the strength of feature associations with both total and motor UPDRS scores.
2. PCA reduced the number of features while preserving ~97.5% of variance in both groups.
3. Linear regression models trained on PCA-transformed data performed better for males than females in terms of both R^2 and error metrics.
4. Predicted vs actual plots confirmed tighter fits for male models.

Analysis:

1. Differences in PCA variance patterns and model performance suggest sex-specific variability in vocal biomarkers.
2. Linear models may not capture the full complexity of the female subgroup; one possible cause could be fewer data points for females (~1900) as compared to males (~4000).
3. Stratifying the dataset by sex was crucial in uncovering these disparities, which would be masked in an aggregate model.

Plan for Additional Analysis:

1. Test interaction terms between features and sex in a combined model.
2. Explore non-linear models such as Random Forest or Gradient Boosting to capture higher-order relation.
3. Compare models using other linear approaches like Lasso and Ridge regression to assess robustness and potential benefits from regularization.
4. Utilize causal inference methodology to evaluate potential cause-effect relationships between vocal biomarkers and disease severity. This would extend the regression framework and help uncover deeper mechanistic insights.