

Motor UPDRS Prediction from Voice Measures with Sex-Specific Effects in Parkinson's Telemonitoring Data

Introduction

Parkinson’s disease is a progressive neurological disorder that affects movement, speech, and many daily motor functions. Because speech can be recorded repeatedly and remotely, voice-based measurements are potentially useful for monitoring symptom severity outside of a clinic. This project uses the Parkinson’s Telemonitoring dataset from the UCI Machine Learning Repository, which contains repeated biomedical voice measurements together with clinician-rated Unified Parkinson’s Disease Rating Scale (UPDRS) outcomes from 42 early-stage Parkinson’s patients. Our response variable is `motor_UPDRS`, and the predictors of primary interest are groups of voice measures related to frequency instability (jitter), amplitude variation (shimmer), noise structure (harmonics-to-noise ratio, NHR), and nonlinear dynamical features (RPDE, DFA, PPE), along with age, sex, and time since recruitment. The research question is:

How well do detailed voice measures predict motor symptom severity in Parkinson’s disease, and do these relationships differ between men and women?

This question is relevant because a useful relationship between voice features and `motor_UPDRS` would support remote symptom monitoring and help identify which voice characteristics are most associated with motor impairment. Linear regression is appropriate because the response is continuous and the project focuses on both *interpretability* and *prediction*. The analysis aims to estimate how average motor symptom severity changes with voice features while also assessing whether sex-specific slopes improve the model. In Parts 1 and 2, we established that a log-transformed response, nonlinear terms, and sex interaction terms improved model diagnostics. In Part 3, we extend the analysis by formally investigating multicollinearity across the voice predictor families, applying all-subsets model selection, and comparing an interpretability-focused model to a purely predictive one.

Methods

The cleaned dataset contained 5,875 observations and 19 predictor variables after removing the subject identifier and the redundant outcome `total_UPDRS`. The analytical workflow is shown in Figure 1 and described in detail below.

Transformation

A raw baseline model showed strongly patterned residuals and non-normal tails, violating the standard regression assumptions. A Box-Cox analysis estimated an optimal power parameter of $\hat{\lambda} \approx 0.707$. Because `motor_UPDRS` is strictly positive throughout the dataset, a log transformation is both theoretically sound and convenient for interpretation. All subsequent models used $\log(\text{motor_UPDRS})$ as the response. Voice predictors that were strictly positive and right-skewed — all jitter and shimmer sub-measures, NHR, HNR, and PPE — were also log-transformed to improve linearity. Squared terms for age, test time, RPDE, DFA, and PPE were added to allow for nonlinear relationships with the response.

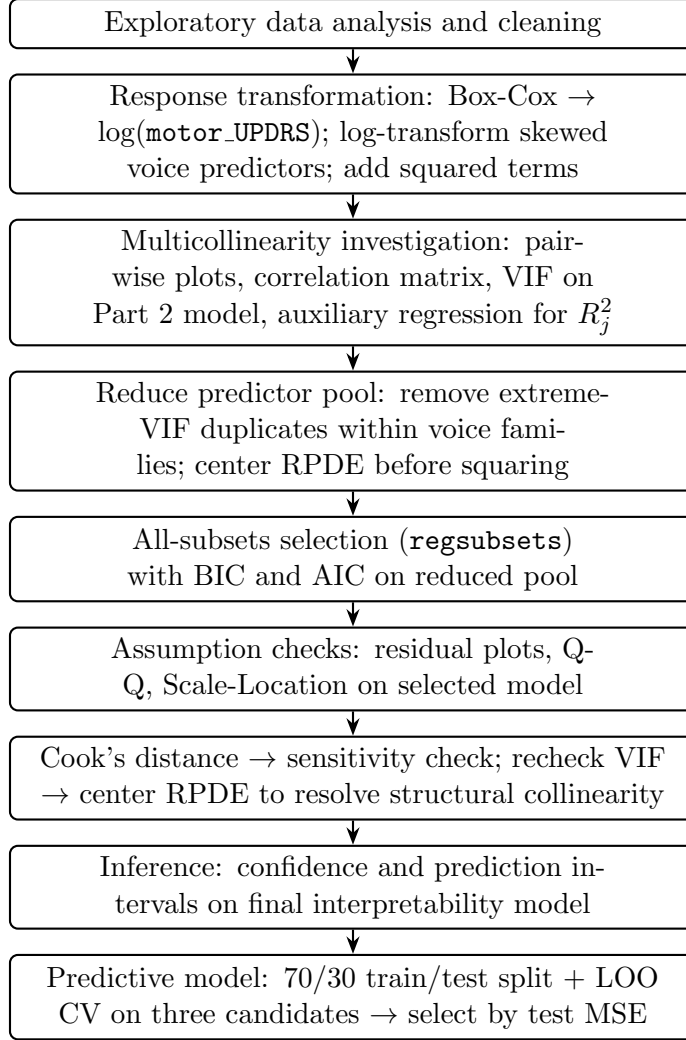


Figure 1: Analytical workflow for building the final interpretability model and the separate predictive model.

Multicollinearity investigation

Before model selection, multicollinearity was investigated through three complementary tools. Pairwise scatterplot matrices were constructed for the jitter family, the shimmer family, and the noise/nonlinear predictor family to visually assess near-linear relationships between sub-measures. The full sample correlation matrix was then computed across all 16 log-transformed voice predictors. Finally, VIFs were computed on the Part 2 final model using the generalized VIF from the `car` package, which correctly handles models with interaction terms. To go beyond reporting VIF values, an auxiliary regression of `log_jitter` on all remaining non-jitter voice predictors was fitted; the resulting R^2 of this auxiliary model directly equals $1 - 1/\text{VIF}$ and quantifies how precisely other predictors can reconstruct `log_jitter`. Based on these findings, a reduced predictor pool was constructed for all-subsets selection by retaining one or two representatives per voice sub-family. RPDE was centered about its sample mean before squaring ($\text{RPDE}_c = \text{RPDE} - \text{RPDE}$) to eliminate the structural collinearity between a variable and its own square.

Model selection

All-subsets selection was applied using `regsubsets()` from the `leaps` package, which searches all 2^p possible submodels and returns the best model of each size by residual sum of squares. AIC and BIC were computed from the resulting RSS values using:

$$\text{AIC} = n \log(\text{RSS}/n) + 2(p + 1), \quad \text{BIC} = n \log(\text{RSS}/n) + \log(n)(p + 1).$$

BIC was used to select the final interpretability model because its stronger penalty on model complexity favours a more parsimonious fit, which supports cleaner coefficient interpretation. The AIC-selected model is also reported for comparison.

Diagnostics and influential observations

After fitting the BIC-selected model, the standard four diagnostic plots, a histogram, and a normal Q-Q plot of standardised residuals were examined. Cook’s distance was computed to identify the most influential observation, and a sensitivity check was performed by removing that observation and refitting. VIFs were then recomputed on the selected model; any remaining structural collinearity was addressed by centering before squaring, as described above. VIF values above 5 were treated as evidence of potential multicollinearity requiring further investigation, consistent with the course convention.

Predictive model

A separate model motivated purely by predictive accuracy was built using a 70/30 train/test split (4,112 training, 1,763 test observations; `set.seed(42)`). Three candidates were trained: (1) the Part 2 nonlinear interaction model, (2) the BIC all-subsets model, and (3) a full additive model with all log-transformed voice predictors and nonlinear terms but no interactions. Test MSE on `log(motor_UPDRS)` was computed for each. LOO CV was also computed using the shortcut formula $\text{MSE}_{\text{LOO}} = n^{-1} \sum_i [\hat{e}_i / (1 - h_{ii})]^2$. The model achieving the lowest test MSE was selected, refitted on the full dataset.

Results

Multicollinearity findings

The pairwise scatterplot matrices revealed near-perfect linear relationships within both voice sub-families. Correlations among the five log-jitter measures ranged from 0.857 to 1.000 (Figure 2), with `log_jitter_rap` and `log_jitter_ddp` perfectly correlated at $r = 1.000$. Within the six log-shimmer measures, correlations ranged from 0.920 to 0.990. Cross-family correlations were elevated but lower: `log_jitter` correlated 0.834 with `log_hnr` and 0.835 with `log_ppe`, while `log_hnr` was negatively correlated with most jitter and shimmer measures (e.g., $r = -0.784$ with `log_shimmer`).

VIFs on the Part 2 final model confirmed the severity of this problem: VIFs for several interaction terms exceeded 10^8 , `log_jitter_rap` and `log_jitter_ddp` exceeded 2×10^5 , and

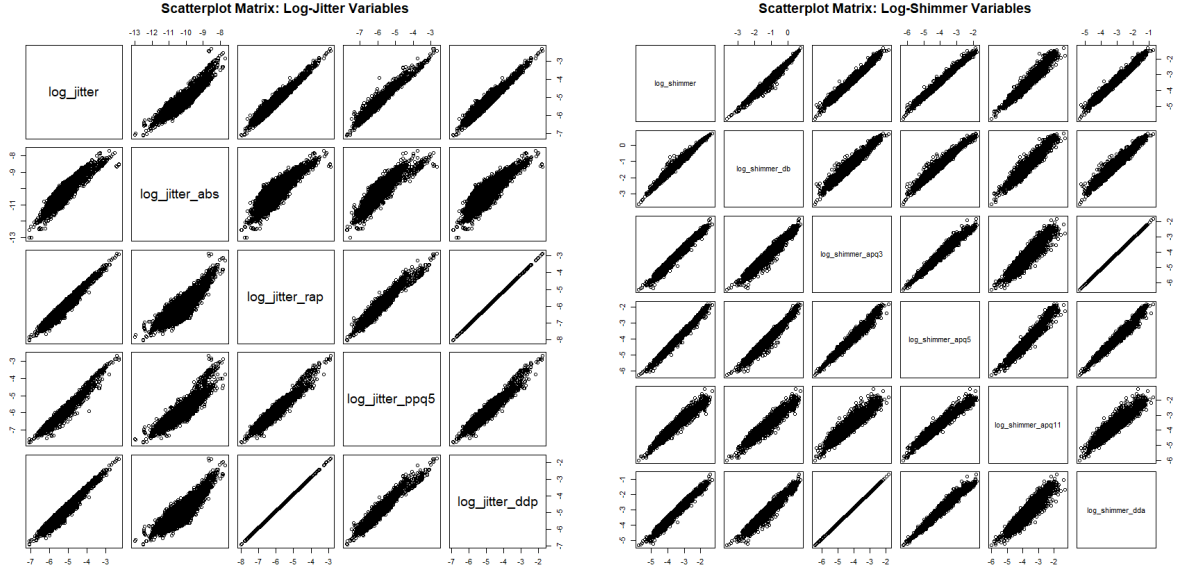


Figure 2: Pairwise scatterplot matrices for the log-jitter family (left) and log-shimmer family (right). Near-linear patterns throughout confirm extreme within-family collinearity, motivating the reduction of each family to a small number of representatives before model selection.

the main effect of `sex` reached 1.5×10^7 due to the large number of interaction terms. The auxiliary regression of `log_jitter` on all non-jitter voice predictors achieved $R^2 = 0.848$, confirming that approximately 85% of the variation in `log_jitter` is linearly explained by the remaining measures. This degree of multicollinearity inflates standard errors substantially and makes marginal t -tests unreliable for individual predictors.

In response, the all-subsets predictor pool was reduced to 17 predictors by retaining `log_jitter` and `log_jitter_abs` (which have distinct acoustic interpretations as percentage versus absolute frequency deviation) as jitter representatives, `log_shimmer` and `log_shimmer_apq5` as shimmer representatives, and all five noise/nonlinear measures. RPDE was centered before squaring, and the remaining within-pool VIFs were all below 20.

All-subsets model selection

All-subsets selection was applied to the 17-predictor reduced pool. BIC selected an 11-predictor model ($\text{BIC} = -10618.68$) and AIC selected a 14-predictor model ($\text{AIC} = -10704.41$). Table 1 shows selected rows from the all-subsets comparison, and Figure 3 shows the BIC plot from `regsubsets`. The BIC-preferred 11-predictor model was chosen as the final interpretability model.

The BIC-selected model included `age`, `test_time`, `sex`, `log_jitter`, `log_jitter_abs`, `log_shimmer_apq5`, `log_hnr`, `RPDE_c`, `DFA`, `log_ppe`, and `RPDE_c_sq`, with no interaction terms. The final interpretability model takes the form:

$$\begin{aligned} \widehat{\log(\text{motor_UPDRS})} = & \hat{\beta}_0 + \hat{\beta}_1 \text{age} + \hat{\beta}_2 \text{test_time} + \hat{\beta}_3 \text{sex} + \hat{\beta}_4 \text{log_jitter} \\ & + \hat{\beta}_5 \text{log_jitter_abs} + \hat{\beta}_6 \text{log_shimmer_apq5} + \hat{\beta}_7 \text{log_hnr} \\ & + \hat{\beta}_8 \text{RPDE_c} + \hat{\beta}_9 \text{DFA} + \hat{\beta}_{10} \text{log_ppe} + \hat{\beta}_{11} \text{RPDE_c_sq}, \end{aligned}$$

Predictors	RSS	Adj. R^2	AIC	BIC
1	1035.75	0.0789	-10192.52	-10179.16
5	969.80	0.1369	-10571.07	-10531.00
9	952.73	0.1516	-10667.37	-10600.59
11	947.00	0.1564	-10698.83	-10618.68
13	945.56	0.1574	-10703.80	-10610.30
14	945.14	0.1576	-10704.41	-10604.23
17	944.58	0.1577	-10701.86	-10581.65

Table 1: Selected rows from the all-subsets model comparison. BIC selects the 11-predictor model; AIC selects the 14-predictor model. Bold rows indicate the optimal model under each criterion.

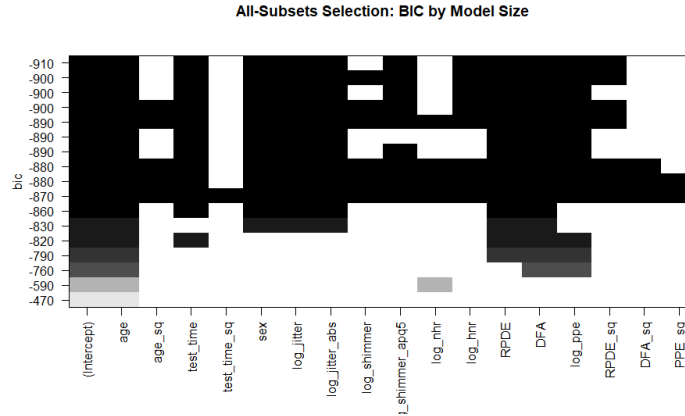


Figure 3: All-subsets BIC plot from `regsubsets()`. Each row represents the best model of a given size; filled cells indicate included predictors. The 11-predictor model (second row from the top) achieves the lowest BIC and is selected as the final interpretability model.

fitted by ordinary least squares on the cleaned dataset of 5,874 observations (after removing the single most influential point). All 11 predictors were statistically significant ($p < 0.001$), with adjusted $R^2 = 0.1564$ and residual standard error = 0.4019 on the log scale.

Post-selection VIF and centering

VIFs on the BIC-selected model before centering showed that $RPDE$ and $RPDE^2$ had VIFs of approximately 68 each — a structural consequence of squaring an uncentered variable. After centering, the VIFs for $RPDE_c$ and $RPDE_c^2$ dropped to 2.1 and 1.1 respectively. The remaining elevated VIFs were $\log_jitter$ (13.1) and $\log_jitter_abs$ (15.7), reflecting genuine acoustic overlap between the two frequency instability measures. Both were retained because they capture distinct aspects of jitter, and removing either would risk omitting a predictor with a meaningful association to motor UPDRS. Centering left the adjusted R^2 unchanged at 0.1564, confirming that only the parameterisation changed.

Diagnostics and sensitivity

Figure 4 shows the diagnostic plots for the centered BIC-selected model. Relative to the raw untransformed baseline, the log transformation substantially reduced residual structure: the

residuals-versus-fitted plot showed no clear non-linear trend, and the Q-Q plot exhibited only mild tail departures. Some heteroscedasticity persisted in the Scale-Location plot, indicating the constant variance assumption was improved but not fully satisfied. Cook's distance identified observation 1881 as the most influential point (Cook's $D = 0.0279$). Removing it and refitting changed adjusted R^2 from 0.1564 to 0.1577 and the residual standard error from 0.4019 to 0.4015 — a negligible shift confirming the conclusions are robust. All inference below uses the sensitivity-checked model.

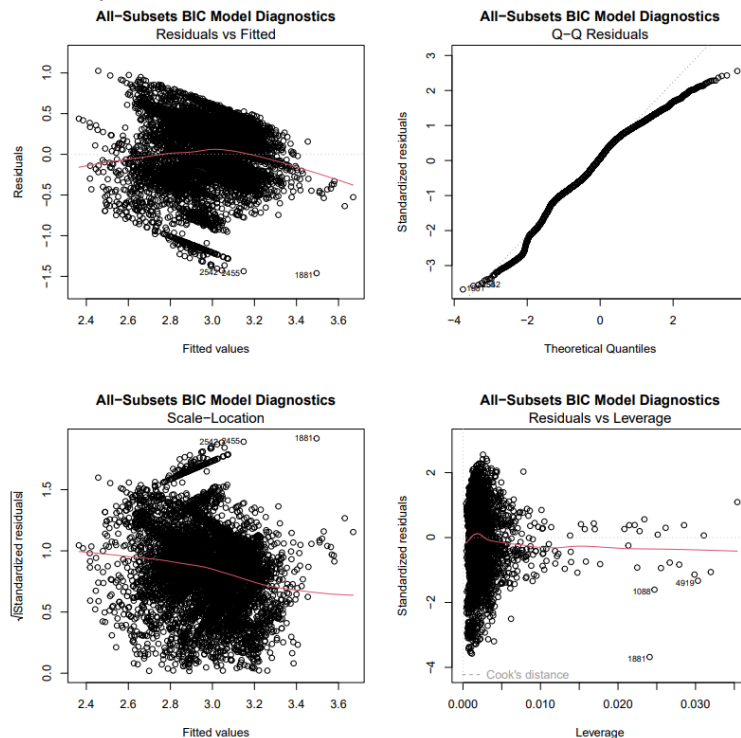


Figure 4: Diagnostic plots for the final BIC-selected interpretability model. The log transformation has substantially reduced residual structure relative to the raw model; mild heteroscedasticity and tail departures persist but do not invalidate the primary conclusions.

Inference

For interpretability, the estimated coefficient for `log_jitter_abs` had a 95% confidence interval of $[-0.381, -0.259]$. Formally, if we were to repeatedly sample from the population and compute a 95% confidence interval from the fitted regression model each time, 95% of those intervals would contain the true population coefficient for `log_jitter_abs`. Because the entire interval is negative, the model provides evidence that, holding all other predictors fixed and for male patients (`sex=0`), increases in `log_jitter_abs` are associated with decreases in the average value of `log(motor_UPDRS)`. The coefficient for `sex` had a 95% confidence interval of $[-0.176, -0.112]$, entirely below zero, indicating that female patients have a lower average `log(motor_UPDRS)` than male patients at the same voice predictor values. The confidence interval for `DFA` was $[-1.020, -0.660]$, providing evidence that greater dynamic complexity in voice is associated with lower average motor UPDRS.

For prediction, voice measures predict individual motor UPDRS with a LOO CV MSE of

0.1618 on the log scale for the interpretability model, corresponding to a root-mean-squared prediction error of approximately 0.402 log-units. On the original scale this translates to individual-level prediction intervals spanning roughly 36 UPDRS points, indicating that voice measures are sufficient for characterising broad population-level symptom trends but not for precise patient-level monitoring. A new observation was constructed using median values of all continuous predictors and `sex`= 0. The 95% confidence interval for the mean `log(motor_UPDRS)` at this predictor combination was [3.008, 3.044], corresponding to approximately [20.25, 20.81] motor UPDRS on the original scale. The 95% prediction interval for a single new observation was [2.235, 3.809], corresponding to approximately [9.34, 45.11] on the original scale. The prediction interval is far wider than the confidence interval, reflecting high individual-level variability in motor UPDRS that the model’s predictors do not capture.

Conclusion and Limitations

The analysis provides a qualified yes to both components of the research question. Voice measures are associated with motor symptom severity: the final BIC-selected model retained 11 predictors across all major voice feature families, all statistically significant, providing evidence that frequency instability, amplitude variation, noise structure, and nonlinear dynamical complexity each carry independent information about motor UPDRS after controlling for the others. The sex coefficient was negative with a confidence interval entirely below zero, indicating that female patients tend to have lower average motor UPDRS than male patients at equivalent voice feature values. This supports the sex-specific component of the research question.

The prediction component is more limited. The model explained approximately 16% of the variation in `log(motor_UPDRS)` (adjusted $R^2 = 0.1577$) and the individual-level prediction interval spanned roughly [9.34, 45.11] on the original scale. The model is therefore more useful for characterising population-level voice–motor associations than for precise patient-level prediction. The findings are broadly consistent with prior Parkinson’s literature, where jitter, shimmer, and HNR are established voice markers of motor deterioration. The negative association between DFA and motor UPDRS (confidence interval [−1.020, −0.660]) suggests that greater signal regularity in voice is associated with higher motor symptom burden, which is also consistent with established findings on nonlinear voice dynamics in Parkinson’s disease.

Several limitations remain. The log transformation improved but did not fully resolve heteroscedasticity and tail departures, so the constant variance assumption is only approximately satisfied. The dataset contains repeated measurements from the same 42 patients, but the modelling framework does not account for within-subject correlation, which may lead to underestimated standard errors and overconfident inference. Moderate VIFs remain for `log_jitter` (13.1) and `log_jitter_abs` (15.7); their individual coefficients should be interpreted cautiously, as the marginal association attributed to one may partly reflect the other. Finally, the analysis is restricted to early-stage Parkinson’s patients and may not generalise to more advanced disease stages.

Comparison to a Predictive Model

Three candidate models were evaluated on test MSE for $\log(\text{motor_UPDRS})$ using a 70/30 train/test split: the Part 2 nonlinear interaction model (P1), the BIC all-subsets model (P2), and a full additive model with no interactions (P3). Test MSEs were 0.14153 (P1), 0.15455 (P2), and 0.15026 (P3). LOO CV confirmed this ranking: 0.15033 (P1), 0.16182 (P2), and 0.15821 (P3). The Part 2 nonlinear interaction model achieved the lowest prediction error on both metrics and was selected as the final predictive model, refitted on the full dataset.

Table 2 compares the two final models. The predictive model uses 45 predictors and achieves an adjusted R^2 of 0.2202 and LOO CV MSE of 0.1503, compared to 11 predictors, adjusted $R^2 = 0.1577$, and LOO CV MSE of 0.1618 for the interpretability model. The improvement in predictive accuracy comes at a cost to interpretability: shared coefficients shift substantially between the two models. For example, the **sex** main effect is -0.144 in the interpretability model but -7.746 in the predictive model, where it is entangled with the large sex interaction structure. Similarly, **log_hnr** shifts from -0.246 to -0.549 and **log_ppe** from 0.140 to 0.417. These differences confirm that the two models serve distinct purposes: the interpretability model provides stable, individually meaningful coefficients for understanding how voice features relate to motor UPDRS, while the predictive model sacrifices interpretability for a lower estimated prediction error.

Metric	Interpretability Model	Predictive Model
Adj. R^2	0.1577	0.2202
RSE (log scale)	0.4015	0.3864
LOO CV MSE	0.1618	0.1503
Number of predictors	11	45
Assumptions required	Yes	No

Table 2: Summary comparison of the final interpretability model (BIC all-subsets, centered RPDE, 11 predictors) and the final predictive model (Part 2 nonlinear interaction, 45 predictors). The predictive model achieves better fit and lower LOO CV MSE at the cost of interpretability.

Statement of AI Use

Claude (Anthropic, claude-sonnet-4-6) was used to assist with improving prose and grammar. ChatGPT (OpenAI) was also used for initial drafting assistance in earlier project parts. All statistical analyses, model fitting, and numerical results were produced exclusively in R by the group.

References

- Anthropic. (2026). *Claude* (claude-sonnet-4-6) [Large language model]. <https://www.anthropic.com>
- Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage. [car package]
- Lumley, T. (2020). *leaps: Regression subset selection* [R package version 3.2]. <https://CRAN.R-project.org/package=leaps>
- OpenAI. (2026). *ChatGPT* [Large language model]. <https://openai.com/chatgpt>
- R Core Team. (2026). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). Accurate telemonitoring of Parkinson’s disease progression by noninvasive speech tests. *IEEE Transactions on Biomedical Engineering*, 57(4), 884–893.
- UCI Machine Learning Repository. (n.d.). *Parkinsons Telemonitoring*. University of California, Irvine, School of Information and Computer Sciences. <https://archive.ics.uci.edu/dataset/189/parkinsons+telemonitoring>