Project and Data Brief

Diagnosing Parkinson's Disease using voice sample data analysis

According to the U.S. National Institute of Neurological Disorders and Stroke, Parkinson's Disease (PD) is a 'movement disorder of the nervous system that gets worse over time'.[1] Common symptoms include tremors, rigidity, bradykinesia, and postural instability. Notable figures who suffered from PD include the legendary boxer Muhammad Ali, the former U.S. president George H.W. Bush, the *Back to the Future* actor Michael J. Fox, the heavy metal legend Ozzy Osbourne, and the late Pope John Paul II.[2]

This Photo by Unknown Author is licensed under CC BY-NC-ND

Sadly, there is currently no known cure for PD. There are also no specific diagnostic tests for the disease. Blood and laboratory tests, as well as brain scans, have been used to rule out other disorders that may cause the symptoms. These diagnostics are invasive with rigorous procedures that may potentially cause further stress for a person living with Parkinson's Disease (PPD).

There is growing interest in developing new, non-invasive methods for diagnosing PD. One possible approach is to analyse a person's speech data obtained from multiple types of sound recordings. PPD usually experiences ***dysphonia***, which leads to reduced loudness, breathiness, roughness, and

[1] National Institute of Neurological Disorders and Stroke (2023). https://www.ninds.nih.gov/health-information/disorders/parkinsons-disease

[2] https://www.parkinson.org/understanding-parkinsons/statistics/notable-figures

exaggerated vocal tremors. These symptoms can be diagnosed in a non-invasive way by analysing various frequency-based characteristics in a person's voice.

This group assessment project focuses on analysing a set of acoustic measurement data extracted from the voice samples of PPD and, where indicated, from those who were healthy. For these subjects, physicians also carried out medical examinations on the PPD and assigned them an appropriate Unified Parkinson's Disease Rating Scale (UPDRS) score to indicate the disease severity and progression.

Datasets

There are two datasets in this project, downloadable from Learnline.

Dataset 1

Filename: **po1_data.txt**

This dataset contains data collected from 40 study subjects, of which 20 were individuals living with Parkinson's Disease (PPD) and the remaining 20 were healthy individuals (non-PPD group).[3] The mean age of the PPD group is 64.86 years old (standard deviation: 8.97). The mean age of the non-PPD group is 62.55 years old (standard deviation: 10.79).

In this study, each subject was asked to record 26 different voice samples. These samples include the recording of saying 3 sustained vowels ("a", "o", "u"), 10 numbers (1 to 10), 9 different words, and 4 rhymed short sentences. This dataset includes all these voice samples but does not indicate the specific recording it represents (i.e. we do not know whether a voice sample comes from recording a sustained vowel or a short sentence). Please note this limitation.

A set of 26 acoustic features were further extracted from each voice sample using a free acoustic analysis software called **Praat**.[4] The variables in this dataset represent these features and are given in Table 1.

Table 1. Description of variables in the po1_data.txt dataset.

Column No.	Measurement category	Description
1	Subject identifier	This number identifies a study subject
2	Jitter	Jitter in %
3	Jitter	Absolute jitter in microseconds
4	Jitter	Jitter as relative amplitude perturbation (r.a.p.)
5	Jitter	Jitter as 5-point period perturbation quotient (p.p.q.5)
6	Jitter	Jitter as average absolute difference of differences between jitter cycles (d.d.p.)
7	Shimmer	Shimmer in %
8	Shimmer	Absolute shimmer in decibels (dB)
9	Shimmer	Shimmer as 3-point amplitude perturbation quotient (a.p.q.3)
10	Shimmer	Shimmer as 5-point amplitude perturbation quotient (a.p.q.5)

[3] Sakar, B.E. *et al*. (2013). https://ieeexplore.ieee.org/abstract/document/6451090

[4] https://www.fon.hum.uva.nl/praat/

11	Shimmer	Shimmer as 11-point amplitude perturbation quotient (a.p.q.11)
12	Shimmer	Shimmer as average absolute differences between consecutive differences between the amplitudes of shimmer cycles (d.d.a.)
13	Harmonicity	Autocorrelation between NHR and HNR
14	Harmonicity	Noise-to-Harmonic ratio (NHR)
15	Harmonicity	Harmonic-to-Noise ratio (HNR)
16	Pitch	Median pitch
17	Pitch	Mean pitch
18	Pitch	Standard deviation of pitch
19	Pitch	Minimum pitch
20	Pitch	Maximum pitch
21	Pulse	Number of pulses
22	Pulse	Number of periods
23	Pulse	Mean period
24	Pulse	Standard deviation of period
25	Voice	Fraction of unvoiced frames
26	Voice	Number of voice breaks
27	Voice	Degree of voice breaks
28	UPDRS	The Unified Parkinson's Disease Rating Scale (UPDRS) score that is assigned to the subject by a physician via a medical examination to determine the severity and progression of Parkinson's disease.
29	PD indicator	Value "1" indicates a subject suffering from PD. Value "0" indicates a healthy subject.

Here are some explanations about the measurement categories.[5] The **jitter** variables measure the variation in the frequency of the sound. The **shimmer** variables measure the variation in the amplitude of the sound. The **harmonicity** variables are related to vocal quality and assess the noise in the sound. The **pitch** variables measure how high or low is the sound based on the frequency of vibration of the sound waves produced. The **pulse** variables measure the glottal pulse, which are variances in voice quality affected by manipulating the vocal cords when speaking. The **voice** variables measure the extent to which a subject has trouble maintaining vocal cord vibration when saying a sustained vowel.

Dataset 2

Filename: **po2_data.csv**

This second dataset contains a range of voice measurements and basic demographic profiles from 42 subjects with early-stage Parkinson's disease. There are no non-PPD subjects in this dataset. Note that these subjects originate from a different study and are different from the subjects described in the first dataset. Each record in this dataset corresponds to a set of features extracted from a voice sample of a subject. The number of voice samples recorded from a subject varies between 101 and 168 voice samples.

Table 2 gives the list of variables in this dataset. The motor UPDRS score could range from 0 to 108, with 0 indicating Parkinson's Disease symptom-free and 108 denoting severe motor impairment. The total UPDRS represents the level of total disability of PPD and ranges from 0 to 176, where 0 denoting a healthy condition and 176 for total disabilities for untreated patients.

[5] Costanzo, P. and Orphanou, K. (2022). https://arxiv.org/abs/2206.03716

Table 2. Description of variables in the po2_data.csv dataset

Column Name	Measurement category	Description
subject#	Subject identifier	This number identifies a study subject
Age	Demography	The subject's age at data collection
sex	Demography	The subject's gender ("0" for male, "1" for female)
test_time	Test detail	The time since subject recruitment into the study. The integer part is the number of days since recruitment.
motor_updrs	UPDRS	The subject's motor UPDRS score assigned by a physician; linearly interpolated.
total_updrs	UPDRS	The subject's total UPDRS score assigned by a physician; linearly interpolated.
jitter(%)	Jitter	Jitter in %
jitter(abs)	Jitter	Absolute jitter in microseconds
jitter(rap)	Jitter	Jitter as relative amplitude perturbation (r.a.p.)
jitter(ppq5)	Jitter	Jitter as 5-point period perturbation quotient (p.p.q.5)
jitter(ddp)	Jitter	Jitter as average absolute difference of differences between jitter cycles (d.d.p.)
shimmer(%)	Shimmer	Shimmer in %
shimmer(abs)	Shimmer	Absolute shimmer in decibels (dB)
shimmer(apq3)	Shimmer	Shimmer as 3-point amplitude perturbation quotient (a.p.q.3)
shimmer(apq5)	Shimmer	Shimmer as 5-point amplitude perturbation quotient (a.p.q.5)
shimmer(apq11)	Shimmer	Shimmer as 11-point amplitude perturbation quotient (a.p.q.11)
shimmer(dda)	Shimmer	Shimmer as average absolute differences between consecutive differences between the amplitudes of shimmer cycles (d.d.a.)
nhr	Harmonicity	Noise-to-Harmonic ratio (NHR)
hnr	Harmonicity	Harmonic-to-Noise ratio (HNR)
rpde	Complex	Recurrence Period Density Entropy
dfa	Complex	Detrended Fluctuation Analysis
ppe	Complex	Pitch Period Entropy

Here are some explanations about the additional measurement categories in this dataset.[6] The **Recurrence Period Density Entropy** variable measures the ability of the vocal folds to sustain simple vibrations. The **Detrended Fluctuation Analysis** describes the extent of turbulent noise in the speech signal. The **Pitch Period Entropy** measures the impaired control of stable pitch during sustained phonation.

Project objectives

Given the speech datasets above, your team is tasked with determining the extent to which the severity of Parkinson's disease can be non-invasively diagnosed (i.e. predicted) using voice sample data.

[6] Tsanas, A. *et al*. (2009). https://www.nature.com/articles/npre.2009.3920.1

Project objective 1

The first project objective requires that your team determine a set of salient variables (features) that could be used to distinguish people with PD from those that are healthy.

This objective must be addressed in Assessment 2. It evaluates your team's ability to apply descriptive and inferential statistical analyses, as well as data wrangling.

You must only use the dataset file **po1_data.txt** dataset for this objective. Please see additional information on Learnline under **Assessment 2**.

Project objective 2

The second project objective requires that your team predict the motor and the total UPDRS scores assigned by a physician to people with Parkinson's Disease.

This objective must be addressed in Assessment 3. It evaluates the overall data science skills that your team learned in this unit, including data exploration, data visualisation methods, and linear regression modelling.

You must at least use the **po2_data.csv** dataset for this objective. Further consideration and experimentation with the **po1_data.txt** dataset to validate your predictive models is essential for scoring any higher grade. Please see additional information on Learnline under **Assessment 3**.

Additional notes

Any creative derivation of new variables from these existing variables or data transformation is permitted. This is known as feature engineering and is a common data science practice. However, you must not alter the original values of these datasets. You must maintain ethical conduct in the course of your data analyses. If in doubt, please consult the Unit Coordinator.