Description and literature:

Parkinson’s disease is a neurodegenerative disorder of central nervous system that causes partial or full loss of motor reflexes, speech, behavior, mental processing, and other vital functions [1]. It is generally observed in elderly people and causes disorders in speech and motor abilities (writing, balance, etc.) of 90% of the patients [2]. Ensuing Alzheimer, PD is the second common neurological health problem in elder ages and it is estimated that nearly 10 million people all around the world and approximately 100k people in Turkey are suffering from this disease [3], [4]. Particularly, PD is generally seen in one out of every hundred people aged over 65. Currently, there is no known cure for the disease [5], [6]. Although, there is significant amount of drug therapies to decrease difficulties caused by the disorder, PD is usually diagnosed and treated using invasive methods [7]. Therefore, this complicates the process of diagnosis and treatment of patients who are grieving from the disease. Our main motivation of working with this dataset is to find a way of identifying patients suffering from Parkinson’s disease with the means of multivariate analysis. We will try to predict the patients who are likely to have the disease by separating the subjects into two groups: those who have Parkinson’s disease and those who don’t.

In this study, using speech data from subjects is expected to help the development of a noninvasive diagnostic. There are important examples of these kinds of Alzheimer and PD studies all around the world [8]. The studies based on the PD focus on symptoms like slowness in movement, poor balance, trembling, or stiffness of some body parts but especially voice problems. The main reason behind the popularity of PD diagnosis from speech impairments is that tele-diagnosis and tele-monitoring systems based on speech signals are low in cost and easy to self-use [6], [8]. Such systems lower the inconvenience and cost of physical visits of PD patients to the medical clinic, enable the early diagnosis of the disease, and also lessen the workload of medical personnel [7], [8]. People with Parkinsonism (PWP) suffer from speech impairments like dysphonia (defective use of the voice), hypophonia (reduced volume), monotone (reduced pitch range), and dysarthria (difficulty with articulation of sounds or syllables). Even though there are many studies aiming at diagnosing and monitoring PD using these impairments, the origin of these studies leans to diagnose basic voice disorders [8]. Therefore, our analysis in this project will be based on voice parameters of the affected. The following section will illustrate a short description of the dataset formation and how we are planning to approach the problem.

Data:

The dataset was created by Athanasios Tsanas and Max Little of the University of Oxford, in collaboration with 10 medical centers in the US and Intel Corporation who developed the tele-monitoring device to record the speech signals. The original study [9] used a range of linear and nonlinear regression methods to predict the clinician's Parkinson's disease symptom score on the UPDRS scale.

This dataset is composed of a range of biomedical voice measurements from 42 people with early-stage Parkinson's disease recruited to a six-month trial of a tele-monitoring device for remote symptom progression monitoring. The recordings were automatically captured in the patient's homes.

Columns in the dataset contain subject number, subject age, subject gender, time interval from baseline recruitment date, motor UPDRS, total UPDRS, and 16 biomedical voice measures. Each row corresponds to one of 5,875 voice recording from these individuals.

The main aim of the data is to predict the motor and total UPDRS scores ('motor_UPDRS' and 'total_UPDRS') from the 16 voice measures.

The data is in ASCII CSV format. The rows of the CSV file contain an instance corresponding to one voice recording. There are around 200 recordings per patient, the subject number of the patient is identified in the first column [10], [11].

ATTRIBUTE INFORMATION:

Subject	Integer that uniquely identifies each subject
Age	Subject age
Sex	Subject gender '0' - male, '1' - female
Test_time	Time since recruitment into the trial. The integer part is the number of days since recruitment
Motor_UPDRS	Clinician's motor UPDRS score, linearly interpolated
Total_UPDRS	Clinician's total UPDRS score, linearly interpolated
Jitter (%) Jitter(Abs) Jitter: RAP Jitter: PPQ5 Jitter: DDP	Several measures of variation in fundamental frequency (Frequency parameters)
Shimmer Shimmer (dB) Shimmer: APQ3 Shimmer: APQ5, Shimmer: APQ11 Shimmer: DDA	Several measures of variation in amplitude (Amplitude parameters)
NHR HNR	Two measures of ratio of noise to tonal components in the voice
RPDE	A nonlinear dynamical complexity measure
DFA	Signal fractal scaling exponent
PPE	A nonlinear measure of fundamental frequency variation

Citations:

[1] J. Jankovic, “Parkinson’s disease: Clinical features and diagnosis,” J. Neurol. Neurosurgery Psychiatry, vol. 79, no. 4, pp. 368–376, 2007.

[2] S. B. O’Sullivan and T. J. Schmitz, “Parkinson disease,” in Physical Rehabilitation, 5th ed. Philadelphia, PA, USA: F. A. Davis Company, 2007, pp. 856–894.

[3] Parkinson Derne˘gi. (2011). [Online]. Available:

http://www. parkinsondernegi.org/Icerik.aspx?Page=parkinsonnedir&ID=5

[4] L. M. de Lau and M. M. Breteler, “Epidemiology of Parkinson’s disease,” Lancet Neurol., vol. 5, no. 6, pp. 525–535, 2006.

[5] N. Singh, V. Pillay, and Y. E. Choonara, “Advances in the treatment of Parkinson’s disease,” Prog. Neurobiol., vol. 81, no. 1, pp. 29–44, 2007.

[6] M. A. Little, P. E. McSharry, E. J. Hunter, J. Spielman, and L. O. Ramig, “Suitability of dysphonia measurements for telemonitoring of Parkinson’s disease,” IEEE Trans. Biomed. Eng., vol. 56, no. 4, pp. 1010–1022, Apr. 2009.

[7] National Collaborating Centre for Chronic Conditions, Parkinson’s Disease, London, U.K.: Royal College of Physicians, 2006.

[8] Betul Erdogdu, SakarMuhammed, Erdem Isenkul, Muhammed Erdem, IsenkulC. Okan, SakarC. and Okan Sakar, “ Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings”, July 2013, IEEE Journal of Biomedical and Health Informatics 17(4):828-834, DOI: 10.1109/JBHI.2013.2245674

[9] Athanasios Tsanas and Max Little, ‘Accurate telemonitoring of Parkinson’s disease symptom severity using nonlinear speech signal processing and statistical machine learning’

[10] Parkinsons Telemonitoring Data Set , Online link: https://archive.ics.uci.edu/ml/machine-learning-databases/parkinsons/telemonitoring/parkinsons_updrs.names

[11] Athanasios Tsanas, Max A. Little, Patrick E. McSharry, Lorraine O. Ramig (2009), 'Accurate telemonitoring of Parkinson.s disease progression by non-invasive speech tests', IEEE Transactions on Biomedical Engineering.

[12] Max A. Little, Patrick E. McSharry, Eric J. Hunter, Lorraine O. Ramig (2009), 'Suitability of dysphonia measurements for telemonitoring of Parkinson's disease', IEEE Transactions on Biomedical Engineering, 56(4):1015-1022