MACHINE LEARNING WITH PYTHON LAB

OPEN ENDED ASSIGNMENT

“Design a machine learning model to predict Parkinson’s disease”

Group Members:

Sr. No	*Name*	*C. No*
1	Gayatri Sadaphal	UEC2021348
2	Manasi Sangamnerkar	UEC2021350
3	Ami Shah	UEC2021354
4	Ishita Shete	UEC2021356

Aim: Develop a machine learning model using Python to classify whether a person has Parkinson's or not using supervised learning.

Introduction:

- Parkinson's disease is a neurodegenerative disorder that affects the nervous system. Early detection of Parkinson's disease is important because it can lead to earlier treatment and management of symptoms. A prediction system for Parkinson's disease can help identify individuals who may be at risk for developing the disease before they start experiencing symptoms.
- This is to help healthcare professionals and researchers involved in the diagnosis and treatment of Parkinson's disease. Early diagnosis leads to early treatment.
- By identifying individuals who are at high risk for developing Parkinson's disease, healthcare professionals can provide targeted interventions and monitoring to help manage symptoms and slow disease progression.

Project Information: The following program demonstrates the use of supervised learning namely classification to help us determine whether a person has Parkinson's disease. This prediction is mainly based on audio frequencies passed and tested on a Parkinson's patient. This is helpful to researchers and healthcare professionals for diagnosing Parkinson's disease, a progressive disorder that affects the nervous system and the parts of the body controlled by the nerves.

Dataset Information:
Source: Kaggle dataset on Parkinson's disease
No. of features: 24
No. of samples: 195
Feature description:

Sr. no	Feature name	Description
1	MDVP: Fo (Hz)	In people with Parkinson's disease, the average vocal fundamental frequency is often lower than in healthy individuals. **This can result in a softer, breathier voice that is more difficult to hear and understand**. The decreased vocal

		fundamental frequency is due to changes in the muscles that control the vocal cords and the respiratory system, which can lead to a decrease in the strength and speed of vocal cord vibrations. This symptom is known as hypophonia and is common in Parkinson's patients
2	MDVP: Fhi (Hz)	In Parkinson's disease, the maximum vocal fundamental frequency can be reduced, which means that **people with Parkinson's may have difficulty reaching high-pitched or loud sounds when speaking or singing**. This can make their voice sound monotone or robotic. The reduction in maximum vocal fundamental frequency is due to the same changes in the vocal cords and respiratory system that cause the decrease in the average vocal fundamental frequency.
3.	MDVP: Flo (Hz)	In Parkinson's disease, the minimum vocal fundamental frequency can also be reduced, which means that **people with Parkinson's may have difficulty reaching low-pitched sounds when speaking or singing.** This can make their voice sound weak or hoarse. The reduction in minimum vocal fundamental frequency is also due to changes in the muscles that control the vocal cords and the respiratory system, which can affect the ability to produce low-pitched sounds.
4	MDVP: Jitter(%)	MDVP: Jitter (%) is a measurement of the variation in the time between each vocal fold vibration during sustained phonation. **In Parkinson's disease, the MDVP: Jitter (%) is often increased**, which means that there is more variability in the vocal fold vibrations during speech. This can result in a jittery or trembling quality to the voice.
5	MDVP: Jitter (Abs)	MDVP: Jitter (Abs) is a measurement of the variation in the period or cycle-to-cycle length of each vocal fold vibration during sustained phonation. **In Parkinson's disease, the MDVP: Jitter (Abs) is often increased**, which means there is more variation in the timing of vocal fold vibrations during speech. This can result in a hoarse or rough-quality of the voice. The increased MDVP: Jitter (Abs) is caused by changes in the nervous system that affect the timing and coordination of the muscles involved in speech production. It is a common characteristic of Parkinson's disease and can be a useful diagnostic tool in conjunction with other symptoms.
6	MDVP: RAP	MDVP: RAP is a measurement of the average rate of change of the sound waveform during the period of vocal fold vibration during sustained phonation. **In Parkinson's disease, the**

		MDVP: RAP is often increased, which means that there is a more rapid change in the sound waveform during the vocal fold vibration period. This can result in a rapid, jerky quality of the voice.
7	MDVP: PPQ	MDVP: PPQ is a measurement of the cycle-to-cycle variation in the sound waveform during sustained phonation. **In Parkinson's disease, the MDVP: PPQ is often increased**, which means there is more variation in the sound waveform during each cycle of vocal fold vibration. This can result in a strained or harsh quality of the voice.
8	Jitter: DDP	Jitter: DDP is a measurement of the difference between the differences in the time between each vocal fold vibration during sustained phonation. I**n Parkinson's disease, the Jitter: DDP is often increased**, which means there is more variation in the timing of vocal fold vibrations during speech. This can result in a rough or unstable quality of the voice.
9	MDVP: Shimmer	MDVP: Shimmer is a measurement of the variation in the amplitude or loudness of the sound waveform during sustained phonation. **In Parkinson's disease, the MDVP: Shimmer is often increased**, which means there is more variation in the loudness of the voice during speech. This can result in a weak or breathy quality of voice.
10	MDVP: Shimmer(dB)	MDVP: Shimmer (dB) is a measurement of the variation in the amplitude or loudness of the sound waveform during sustained phonation, expressed in decibels. **In Parkinson's disease, the MDVP: Shimmer (dB) is often increased**, which means there is more variation in the loudness of the voice during the speech, as measured in decibels. This can result in a weak or breathy quality of voice.
11	Shimmer: APQ3	Shimmer: APQ3 is a measurement of the short-term variation in the amplitude or loudness of the sound waveform during sustained phonation. **In Parkinson's disease, the Shimmer: APQ3 is often increased,** which means there is more short-term variation in the loudness of the voice during speech. This can result in a weak or breathy quality to the voice. .
12	Shimmer: APQ5	Shimmer: APQ5 is a measurement of the long-term variation in the amplitude or loudness of the sound waveform during sustained phonation. **In Parkinson's disease, the Shimmer: APQ5 is often increased**, which means there is more long-term variation in the loudness of the voice during speech. This can result in a weak or breathy quality to the voice.

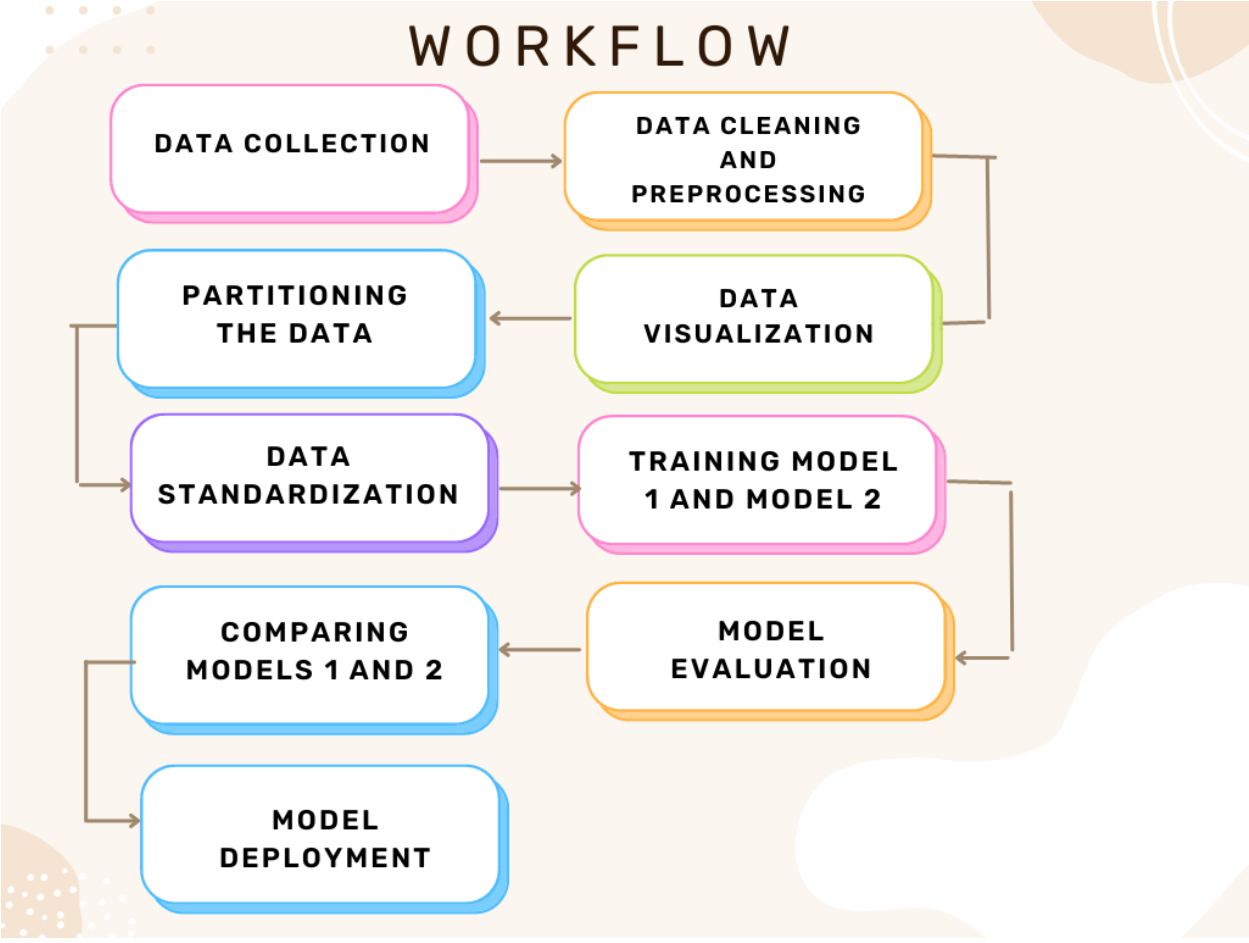
13	MDVP: APQ	MDVP: APQ is a measurement of the degree of irregularity or complexity of the sound waveform during sustained phonation. **In Parkinson's disease, the MDVP: APQ is often increased**, which means that there is more irregularity or complexity in the sound waveform during speech. This can result in a rough or strained quality of the voice.
14	Shimmer: DDA	Shimmer: DDA is a measurement of the difference between the absolute values of consecutive peaks and valleys in the amplitude or loudness of the sound waveform during sustained phonation. **In Parkinson's disease, the Shimmer: DDA is often increased**, which means there is more variation in the loudness of the voice during speech. This can result in a weak or breathy quality of voice.
15	NHR	NHR stands for "Noise-to-Harmonics Ratio," which is a measure of the ratio of noise to tonal components in the voice. **In Parkinson's disease, the NHR is often increased**, which means that there is more noise in the voice compared to the harmonics or tonal components. This can result in a harsh or hoarse quality of the voice.
16	HNR	HNR stands for "Harmonics-to-Noise Ratio," which is a measure of the ratio of harmonics or tonal components to noise in the voice. **In Parkinson's disease, the HNR is often decreased**, which means that there are fewer harmonics or tonal components in the voice compared to the noise. This can result in a breathy or weak quality of the voice.
17	RPDE	RPDE stands for "Recurrence Period Density Entropy," which is a measure of the degree of regularity or complexity of the voice signal during sustained phonation. **In Parkinson's disease, the RPDE is often increased**, which means that there is more irregularity or complexity in the voice signal during speech. This can result in a rough or strained quality of the voice.
18	D2	D2, also known as Correlation Dimension or Correlation Integral, is a nonlinear measure that can be used to analyze the speech signal in Parkinson's disease. D2 measures the degree of complexity or irregularity in a time series, by examining how the points in a phase space are distributed. **In Parkinson's disease, D2 is often increased**, indicating a more complex or irregular distribution of the F0 values in the speech signal. This increase in D2 may reflect the loss of control and coordination in the muscles involved in

		speech production, resulting in a more erratic and unpredictable F0 contour.
19	DFA	DFA, or Detrended Fluctuation Analysis, measures the degree of long-term correlation or persistence in a time series, by examining the scaling properties of the fluctuations in the signal. **In Parkinson's disease, DFA is often decreased**, indicating a reduction in long-term correlation or persistence in the speech signal. This decrease in DFA may reflect the loss of control and coordination in the muscles involved in speech production, resulting in a more random and unpredictable F0 contour.
19	Spread 1	Spread 1 is a nonlinear measure of fundamental frequency variation that can be used to analyze the speech signal in Parkinson's disease. Spread 1 is a measure of the dispersion or spread of the distribution of the F0 values around the mean. **In Parkinson's disease, Spread 1 is often decreased**, which indicates a narrower distribution of F0 values around the mean. This reduction in Spread 1 may reflect the loss of flexibility and range of movement in the muscles involved in speech production, resulting in a more monotonic and less varied F0 contour. Like other nonlinear measures of F0 variation, Spread 1 can provide additional insights into the changes in speech production in Parkinson's disease beyond the traditional measures of jitter and shimmer.
20	Spread 2	It is a measure of the regularity or predictability of the distribution of F0 values around the mean. **In Parkinson's disease, Spread 2 is often increased**, which indicates a more irregular or unpredictable distribution of F0 values around the mean. This increased Spread 2 may reflect the loss of control and coordination in the muscles involved in speech production, resulting in a more chaotic and unstable F0 contour.
21	PPE	PPE, or Pitch Period Entropy, is a nonlinear measure that can be used to analyze the speech signal in Parkinson's disease. PPE measures the degree of variability or unpredictability in the pitch period of the speech signal. **In Parkinson's disease, PPE is often increased**, indicating a more irregular or unpredictable pitch period. This increase in PPE may reflect the loss of control and coordination in the muscles involved in speech

		production, resulting in a more disordered and less stable F0 contour.
22	Status	Health status of the subject: one, Parkinson's; zero, healthy

- ☐ Five measures of variation in fundamental frequency: MDVP Jitter(%), MDVP Jitter (Abs), MDVP RAP, Jitter DDP, MDVP PPQ
- ☐ Six measures of variation in amplitude: MDVP Shimmer, MDVP Shimmer(dB), Shimmer APQ3, Shimmer APQ5, MDVP APQ, Shimmer DDA
- ☐ Two measures of the ratio of noise to tonal components in the voice: NHR, HNR
- ☐ Two nonlinear dynamical complexity measure: RPDE, D2
- ☐ Three nonlinear measures of fundamental frequency variation: Spread 1, Spread 2, PPE

Workflow:



Designing the ML model:

1] Data collection: The dataset of Parkinson's prediction model has been collected from Kaggle. It has 24 features and 195 samples. The dataset is based on the voice modulation of a person and the study of frequency distribution.

2] Data cleaning and preprocessing:

Data Cleaning:

- The data is grouped according to the outcomes ("status"). Here, 1 indicated that the person is Parkinson's positive and 0 indicated the person is Parkinson's negative.
- Then the data is checked for missing values using the ".isnull().sum()" function to group all the null values.

Data Preprocessing

- Specific columns such as "name" is removed because it isn't involved in predicting whether a person has Parkinson's or not.
- Partitioning the data into training data and testing data.

3] Data Visualization: Various features of the dataset are represented through the scatter plot to obtain the outliers. However, there are different outliers for different features so none of the outliers are removed.

4] Splitting the data: The data is split into training and testing data by a ratio of 80:20.

5] Data Standardization: we want to transform our dataset so that each feature (or column) has a similar scale and range of values.

6] Training the models:

- Model 1:Support Vector Machine
- Model 2: Naive Bayes

7] Model Evaluation: M

❖ Model 1: SVM

Accuracy score of test data: 0.8717

Confusion Matrix:

5	3
2	29

- Classification report:

	Precision	Recall	F1-score	support
0	0.71	0.62	0.67	8
1	0.91	0.94	0.92	31
accuracy			0.87	39
Macro avg	0.81	0.78	0.79	39
Weighted avg	0.87	0.87	0.87	39

❖ **Model 2:**

Accuracy score of test data: 0.6153

Confusion Matrix:

8	0
15	16

- Classification report:

	Precision	Recall	F1-score	support
0	0.35	1.00	0.52	8
1	1.00	0.52	0.68	31
accuracy			0.62	39
Macro avg	0.67	0.76	0.68	39
Weighted avg	0.87	0.62	0.65	39

8] Selecting the better model: Since the recall for class ‘1’ of the model 1 is better than compared to class ‘1’ for model 2, SVM model is selected. Recall is an important metric in evaluating the performance of a model, particularly in cases where correctly identifying positive cases is crucial, such as in medical diagnosis.

9] Model Deployment: The model is then given a set of test values on which it decides whether a person has Parkinson’s disease or not

Results:
Since the SVM model is better, therefore, the metrics of the same are considered:

	Precision	Recall	F1-score	support
0	0.71	0.62	0.67	8
1	0.91	0.94	0.92	31
accuracy			0.87	39
Macro avg	0.81	0.78	0.79	39
Weighted avg	0.87	0.87	0.87	39

From above we can observe that the accuracy is 0.87, recall for class 1 is 0.94

Conclusion:

A recall score, also known as sensitivity or true positive rate, is a performance metric that measures the proportion of actual positives that are correctly identified by a classification model.

Here, 2 models were compared on the basis of their class '1' recall and we found out that the SVM model works best as it is useful for high-dimensional datasets.

- SVM works well with high-dimensional data, it can handle a large number of features, making it useful for datasets with many variables. Naive Bayes, on the other hand, may not perform as well with high-dimensional data.

- SVM can handle non-linear data, it can use kernel functions to transform the data into a higher-dimensional space where it may be easier to separate the classes. This allows SVM to handle non-linear data better than Naive Bayes.

The whole model is useful for detection whether a person has Parkinson's disease or not.