Project Scope

Problem Statement: The project focuses on predicting Parkinson's disease status using a dataset containing various clinical and demographic features. The primary objective is to develop a reliable machine learning model to assist in early detection, which can lead to timely intervention and better management of the disease.

Challenges: One of the primary challenges was dealing with an imbalanced dataset, as this can lead to a biased model favoring the majority class. Additionally, selecting the most relevant features from the dataset and fine-tuning the model to achieve optimal performance required careful attention to detail.

My Role and Solution

Role: Sole Data Scientist

Solution Approach: I worked alone on the project, overseeing data exploration, preprocessing, and model development. Using Python, I implemented a decision tree classifier as our primary model. The approach involved iterating through feature selection, model training, and hyperparameter tuning using GridSearchCV to refine the model's performance. Additionally, I used various data visualization techniques to better understand feature importance and model predictions.

My Work Process

Involvement: I was responsible for every stage of the data pipeline:

1. Imports and Setup: Imported essential libraries such as pandas, scikit-learn, and matplotlib to handle data manipulation, model training, and visualization.
2. Data Loading: Loaded the dataset using pandas and examined the structure of the data to understand its composition.
3. Data Preparation: This involved dropping missing values and irrelevant columns (e.g., 'name'), followed by one-hot encoding categorical variables to make the data suitable for model training.
4. Modeling: A decision tree classifier was chosen for its interpretability. I performed hyperparameter tuning with GridSearchCV to identify the best parameters for the model.
5. Evaluation: The model was evaluated using metrics such as accuracy, precision, recall, and F1-score. A confusion matrix was also plotted to visualize the model's performance on the test set.

Technical Contributions:

- Feature Importance: I plotted feature importance to determine which features had the most influence on the model's predictions.
- Confusion Matrix: Generated a confusion matrix heatmap to provide insights into the model's classification performance.

Outcome and Results

Outcome: The final decision tree model demonstrated strong predictive capabilities, with well-balanced performance metrics across both classes. The visualization of feature importance also provided meaningful insights, highlighting key features that could be used for further medical investigations.

Results: The project concluded with a robust model that can be applied to real-world data for predicting Parkinson's disease. The insights gained from feature importance and model evaluation offer significant potential for early diagnosis and targeted treatments.

Performance metrics

Metric	Train	Test
Accuracy	1.0000	0.9474
Precision	1.0000	0.9545
Recall	1.0000	0.9767
F1 Score	1.0000	0.9655

	0	1	2	3	4	5	6	7	8	9	...	180	181	182	183
MDVP:Fo(Hz)	119.992	122.4	116.682	116.676	116.014	120.267	107.332	95.73	95.056	88.333	...	116.286	116.556	116.342	114.563
MDVP:Fhi(Hz)	157.302	148.65	131.111	137.871	141.781	137.244	113.84	132.068	120.103	112.24	...	177.291	592.03	581.289	119.167
MDVP:Flo(Hz)	74.997	113.819	111.555	111.366	110.655	114.82	104.315	91.754	91.226	84.072	...	96.983	86.228	94.246	86.647
MDVP:Jitter(%)	0.00784	0.00968	0.0105	0.00997	0.01284	0.00333	0.0029	0.00551	0.00532	0.00505	...	0.00314	0.00496	0.00267	0.00327
MDVP:Jitter(Abs)	0.00007	0.00008	0.00009	0.00009	0.00011	0.00003	0.00003	0.00006	0.00006	0.00006	...	0.00003	0.00004	0.00002	0.00003
MDVP:RAP	0.0037	0.00465	0.00544	0.00502	0.00655	0.00155	0.00144	0.00293	0.00268	0.00254	...	0.00134	0.00254	0.00115	0.00146
MDVP:PPQ	0.00554	0.00696	0.00781	0.00698	0.00908	0.00202	0.00182	0.00332	0.00332	0.0033	...	0.00192	0.00263	0.00148	0.00184
Jitter:DDP	0.01109	0.01394	0.01633	0.01505	0.01966	0.00466	0.00431	0.0088	0.00803	0.00763	...	0.00403	0.00762	0.00345	0.00439
MDVP:Shimmer	0.04374	0.06134	0.05233	0.05492	0.06425	0.01608	0.01567	0.02093	0.02838	0.02143	...	0.01564	0.0166	0.013	0.01185
MDVP:Shimmer(dB)	0.426	0.626	0.482	0.517	0.584	0.14	0.134	0.191	0.255	0.197	...	0.136	0.154	0.117	0.106
Shimmer:APQ3	0.02182	0.03134	0.02757	0.02924	0.0349	0.00779	0.00829	0.01073	0.01441	0.01079	...	0.00667	0.0082	0.00631	0.00557
Shimmer:APQ5	0.0313	0.04518	0.03858	0.04005	0.04825	0.00937	0.00946	0.01277	0.01725	0.01342	...	0.0099	0.00972	0.00789	0.00721
MDVP:APQ	0.02971	0.04368	0.0359	0.03772	0.04465	0.01351	0.01256	0.01717	0.02444	0.01892	...	0.01691	0.01491	0.01144	0.01095
Shimmer:DDA	0.06545	0.09403	0.0827	0.08771	0.1047	0.02337	0.02487	0.03218	0.04324	0.03237	...	0.02001	0.0246	0.01892	0.01672

Dataset used

Confusion matrix