Analysis of the Changing Nature of Data Science in the Time of Generative AI through examination of of the Parkinson's Study of Data versus Voice Data and

The problem statement I initially had was to take a set of data describing an individual and determine whether or not, based on the data recorded, this individual has Parkinson's disease. While my capstone focused on this analysis for this class, I have a much more important conclusion to express … This is, during the eleven months that I have been in this certificate program, the world of data science and machine learning has changed so significantly that where we began with and where we are today are no longer even comparable.

During the initial phase of my interaction with the subject matter in this course, I used Jupyter notebook extensively, meticulously going through each line of code, running it, correcting errors. This is the same approach to programming that I had taken for the last 30+ years, starting in 1997 with VBA, continuing into programming in Python in 2021. I am familiar with this approach and it helps me learn. In fact, prior to GenAI and ChatGPT, this may have been the approach many took. In that data science and AI/ML taught in this course relies upon a solid foundation and approach to programming, this makes sense. The period for which I applied this approach was October through December 2024.

The course introduced me to Google Collab, which includes Gemini. Along with the occasionally use of ChatGPT, I found it more helpful, faster, more expedient, and more efficient to interact with Gemini or ChatGPT to debug my code, help improve it, comment on it, recommend solutions, etc. This accelerated my learning, made me a better programmer, and led to better, more impactful results, faster. My initial data analysis for the capstone was done in this manner. However, may times the code assistance created problematic, buggy, or unusable code. In general it was great, but still had a very heavy reliance on my input and development. This period lasted from January 2025 to May 2025.

For personal reasons, I had to take a hiatus in the certificate program, and did not re-engage until late August 2025. The world changed during this hiatus. At my office, a coworker recommended I try Anthropic Claude. Using Claude, and presenting it with the problem statement and programming approach, that I had initially identified for the capstone, without less than 3 minutes, Claude generated one single cell of under 100 lines of code, written with ZERO errors, which executed, flawlessly in Jupyter notebook on my Mac laptop. I ended up interacting with Claude over the course of a number of hours to simplify the code, and added in items I wanted, like feature engineering, but my interaction with the code was reduced to the point where Claude is now like a calculator or even a spreadsheet. All of the hand written formulas and calculations I used to do for financial analysis were automated away for work with spreadsheets…. Now all of the algorithms are automated away with this GenAI code assistant / code generator.

This means that, if one understands the business context of problem, identifies the data set, and understands the underlying implications of the choices made on the models, the tuning, hyper tuning, etc., the time barrier has been reduced or removed to make data science available to everyone with the subject matter expertise, and no longer much, if any reliance on programming

ability. This will hugely impact the field of data science, ML and AI, much greater than anyone would have thought in October 2024, and in fact these ripples are being felt as shockwaves through the industry. Programming has been automated, and with it, the knowledge of how to interact with data is being unleashed. The new emphasis will be on, do people understand how to interpret,analyze, review, make reasonable, accurate conclusions, and come up with an appropriate and beneficial plan of action based on it. I believe this is the most important conclusion I can have, versus whether or not a model can be used to identify Parkinsons… The answer to the second question is yes, absolutely, and the power to do so is now in the hands of everyone, if they desire to learn to do so, and not only for his problem space, but all problems.

In the Pentagon, where I used to work, there was a statement outside the office the the Dept. of Defense Chief Information Officer's door. It said, "In God We Trust, All Others Must Bring Data."

How true this is…

Now, what follows is the analysis of the data.

My interest in this area was due to the fact that my mother started exhibiting symptoms of cognitive decline. After much testing, and many visits to various health care providers and medical specialists, the diagnosis was inconclusive, as her prescribed diagnosis was "atypical" Parkinson's. Atypical Parkinson's essentially indicates a series of observed symptoms which, when taken together, have the appearance of Parkinson's, but the root cause of these symptoms is not determined. There are many data sets using brain imaging, handwriting, demographic details, lifestyle details and clinical measurements which can be used to determine if an individual has Parkinson's.

Using voice data, and PCA analysis with a linear regression model, the model had an accuracy score of 0.9365 and 0 false negatives, with 100% recall. From a medical perspective however, relying on voice data may be problematic, because of the need for specialized equipment and a standardized environment to capture the voice data. For this reason, I used a set of disease data which included many different medical observations as follows:

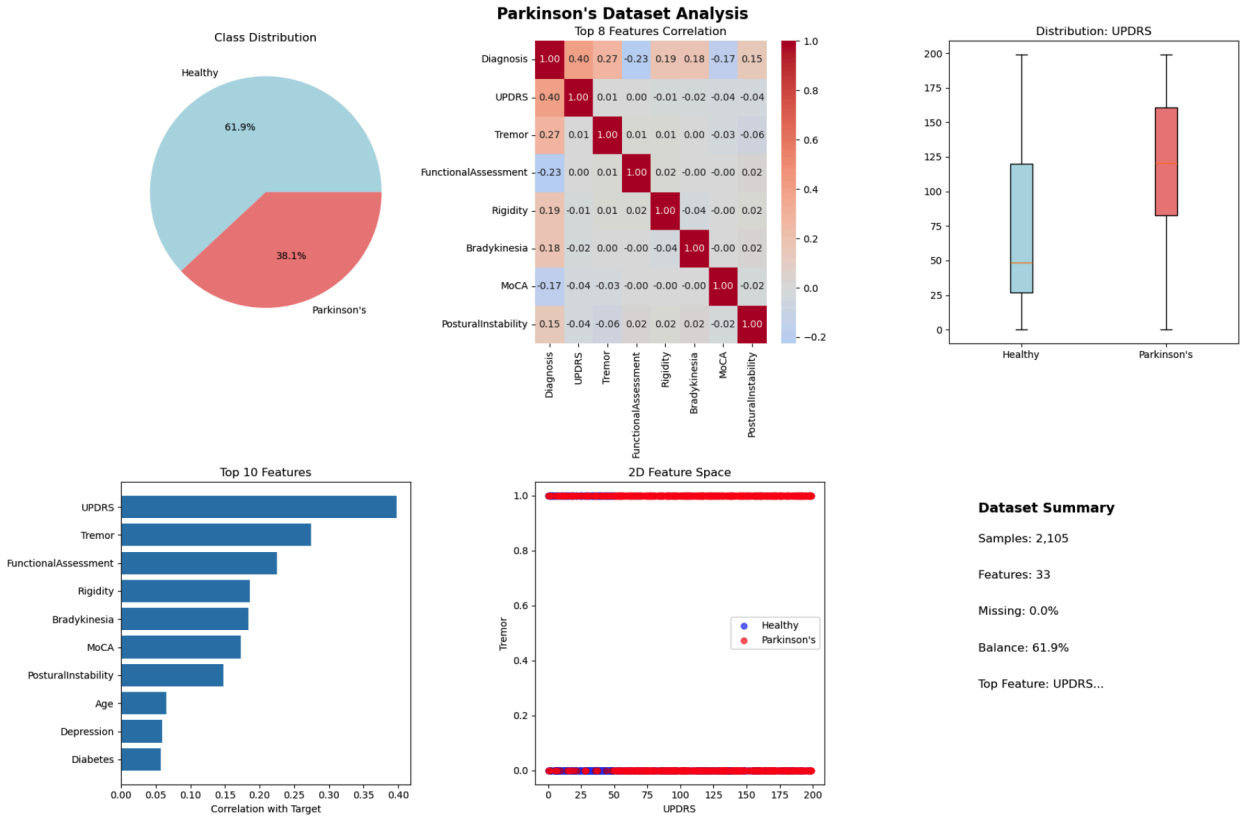
🔝 TOP 15 FEATURES BY CORRELATION STRENGTH:

Rank	Feature	Correlation	Strength
1	UPDRS	0.398	Moderate
2	Tremor	0.274	Weak
3	FunctionalAssessment	-0.225	Weak
4	Rigidity	0.186	Weak
5	Bradykinesia	0.184	Weak
6	MoCA	-0.173	Weak
7	PosturalInstability	0.148	Weak
8	Age	0.065	Very Weak
9	Depression	0.059	Very Weak
10	Diabetes	0.057	Very Weak
11	PatientID	-0.044	Very Weak

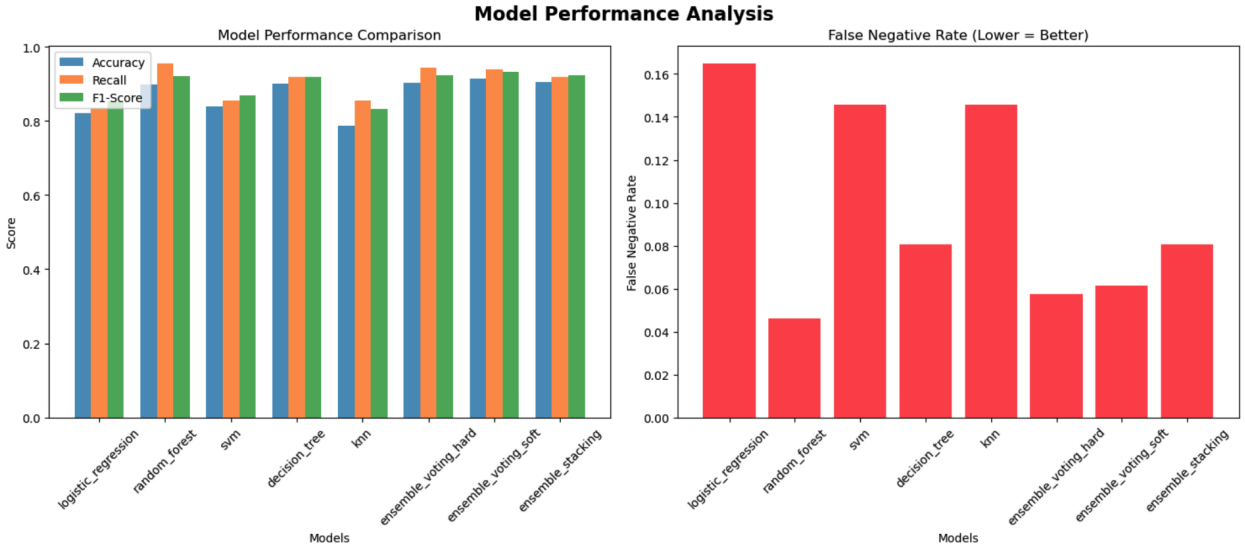
12	SleepQuality	-0.043	Very Weak
13	AlcoholConsumption	0.037	Very Weak
14	BMI	0.030	Very Weak
15	DiastolicBP	-0.029	Very Weak

Using this new data set, and applying an ensemble technique, incorporating feature engineering to encode categorical labels, and grid search to identify the best parameters and model, the new data set was able to show an accuracy score of 89.9% and a recall of 95.4%. Above my target value of 85%, but not as performant as the original analysis on voice data alone. Please refer to the initial report.

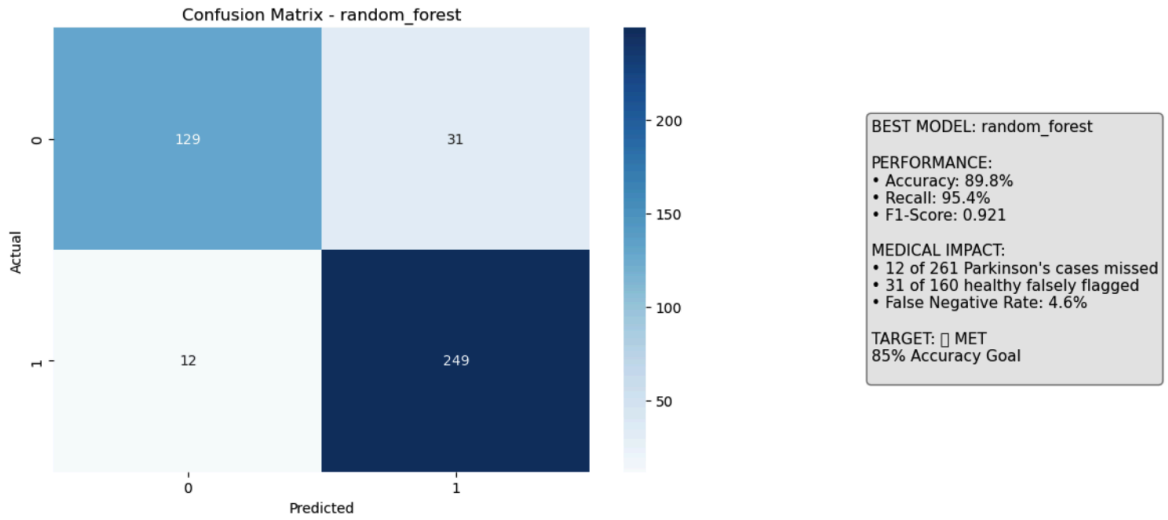
The following images illustrate the Parkinson's Disease Data Set.



The following images show the model performance, where random forest is the best model.



The following images confusion matrix and best model information shows the details.




```
🎯 FINAL SUMMARY:
Best model: random_forest
Accuracy: 89.8%
Medical safety: 12 Parkinson's cases missed out of 261
85% target: Achieved
```

sadfasdf