BSTA 103 Exam

This is an individual exam, and no collaboration or communication with others is allowed. Any attempt to share, seek assistance, or engage in unauthorized communication will be considered a breach of academic integrity. However, you are encouraged to contact the instructor if you have any questions or need clarification.

The problem

The dataset, available on the course site, comprises handwriting data from 174 participants who completed 25 tasks as described in the article titled “Diagnosing Alzheimer?s disease from online handwriting: A novel dataset and performance benchmarking" by Cilia et al. (2022). The primary goal is to distinguish Alzheimer's disease patients from healthy individuals. These tasks included both memorization, drawing and dictation exercises. In the memory tasks, participants were tasked with reproducing previously memorized words or words associated with specific objects. The dictation tasks aimed to understand how handwriting characteristics change when participants use their working memory. Specifically, participants were asked to memorize and rewrite the words such as “telefono", “cane", and “negozio". From the raw data, which includes $(x, y)-$coordinates, pressure, and timestamps, 18 features were extracted. Below is an explanation of what each column represents:

Total Time (TT)	Total time spent to perform the entire task.
Air Time (AT)	Time spent to perform in-air movements.
Paper Time (PT)	Time spent to perform on-paper movements.
Mean Speed on-paper (MSP)	Average speed of on-paper movements.
Mean Speed in-air (MSA)	Average speed of in-air movements.
Mean Acceleration on-paper (MAP)	Average acceleration of on-paper movements
Mean Acceleration in-air (MAA)	Average acceleration of in-air movements.
Mean Jerk on-paper (MJP)	Average jerk of on-paper movements.
Mean Jerk in-air (MJA)	Average jerk of in-air movements.
Pressure Mean (PM)	Average of the pressure levels exerted by the pen tip.
Pressure Var (PV)	Variance of the pressure levels exerted by the pen-tip.
GMRT on-paper (GMRTP)	Generalization of the Mean Relative Tremor (MRT) computed for on-paper movements.
GMRT in-air (GMRTA)	Generalization of the Mean Relative Tremor computed on in air movements.
Mean GMRT (GMRT)	Average of GMRTP and GMRTA.
Pendowns Number (PWN)	Counts the total number of pendowns recorded during the execution of the entire task (e.g., a continuous uninterrupted line present a pendowns number equal to 1)
Max X Extension (XE)	Maximum extension recorded along the X-axis.
Max Y Extension (YE)	Maximum extension recorded along the Y-axis.
Dispersion Index (DI)	This measures how the handwritten trace is “dispersed" on the entire piece of paper; in other words, it measures how much of the sheet is covered
Class	**P**: Alzheimer's disease patients. **H**: healthy people

Please extract data of the 16^{th} task to answer the following questions. When you load the data, set the first column as the index. You may read the paper to gain more insight into the data.

1. Make an appropriate plot to determine whether the data is balance or not.
2. Build and evaluate ML predictive models using the following algorithms:

   - Logistic regression
   - Decision Tree

   by splitting your extracted data with 75% for training your models. Set the random generator seed to 14 and rescale your data using the $\min - \max$ scaler. Choose hyperparameters within or equal to the ranges specified in Table 4 of the paper.
3. Present a table of an appropriate metric that compares the performance of the two models and interpret them. Your table should be only three columns (thus algorithm, Metric and interpretation).
4. Which of your two models is performing better?
5. Identify the mean accuracy of the two models from Table 7 of the paper and then compare them with your models.

Submission

Attach your python program (**.ipynb**).