Scientific Programming with Python - Exercise 3

Attached to this exercise is a CSV dataset, "penguins.csv". The data set contains attributes of penguins from different species. You are tasked with creating and visualizing a Gaussian Naïve Bayes classifier model for this data set. Your submission may be in pptx/pdf/docx format (your choice) alongside the course code, or a complete Jupyter notebook.

Task 1 – predict species

1. Select the 2 features which allow for the most accurate 2-feature GNB classifier. Explain your selection.
2. Train your model using 80% of the data set as your training set.
3. Use a filled contour plot to show the decision distribution of your model (limit your plot axes to the actual data boundaries +-1).
4. Overlay a scatter plot containing **only the points from the original data set that would have had incorrect predictions given this model** above the filled contour plot.
5. Include the classification report for your model with regards to the test set.

Task 2

1. Create a new column named "class" in the data set which is a combination of the "sex" and "species" category, e.g. Male Adelie, Female Chinstrap, etc…
2. Repeat task 1 using "class" as your target class.

Grading:

40 points for correctness.

40 points for esthetics (visualization)

20 points for code cleanliness and readability

Task 5 – Bonus Task!

This task is worth 25 bonus points and can bring your exercise grade to above 100.

HOWEVER, these bonus points will be awarded to up to 3 student pairs. When grading submissions, if we find that more than 3 pairs have successfully answered this bonus question, points will go to the earliest submissions (based on the last submission time of each pair). Do not share your answer with other students!

1. In Task 4 question 3 you were asked to find the feature on which the transformation of price into price_2 was made. You must now find that exact manner in which this transformation was done. Analyze the data and present the algorithm used when transforming price into price_2.