Task 2A: Critical Thinking Assessment: [/ 2P]

The proposed solution to the specified problem statement has many inherent flaws that can be seen from these visualizations (or flaws that will appear on choosing a different dataset for a similar problem statement). **Please mention ONE of the main flaws in one of the two proposed Visualization.** This task is meant to assess critical thinking which is necessary in identifying research gaps/problems with existing solutions.

→ Flaws :- Split Attention Effect, Overloading of color Encoding, Class Imbalance Representation.

Out of these, the main flaws can be seen in task 1b i.e Class Imbalance Representation in the dataset (7000 dogs images vs 3000 cat images) which is a crucial issues inadequately addressed by the bar chart in Task 1b. The result can mislead viewers into thinking that the classifiers are more effective on the dogs class due to the higher number of dogs images, rather than because of true performance differences. This distortion can lead to incorrect conclusions about the classifiers effectiveness. too.

Task 2B: Literature Search: [/ 1P]

Kindly provide ONE literature reference (paper) that pertains to a comparable problem statement and employs information visualization or visual analytics techniques to address the challenge.

Paper 1 :- "Visual Analytics for Machine learning: A data Perspective Survey"

Authors :- Junpeng Wang, Shixia Liu, Wei Zhang.

→ This paper gives a wide range of visualization techniques that could be applied to improve the bar chart's ability to convey class imbalances in datasets

Paper 2 :- "Tackling Class Imbalance in Computer Vision: A Contemporary Review".

Authors :- Manisha saini, seba Susan.

→ This paper reviews the latest techniques used in computer vision to handle class imbalance, including how visualization can be leveraged to better understand and correct the bias introduced by imbalanced datasets.

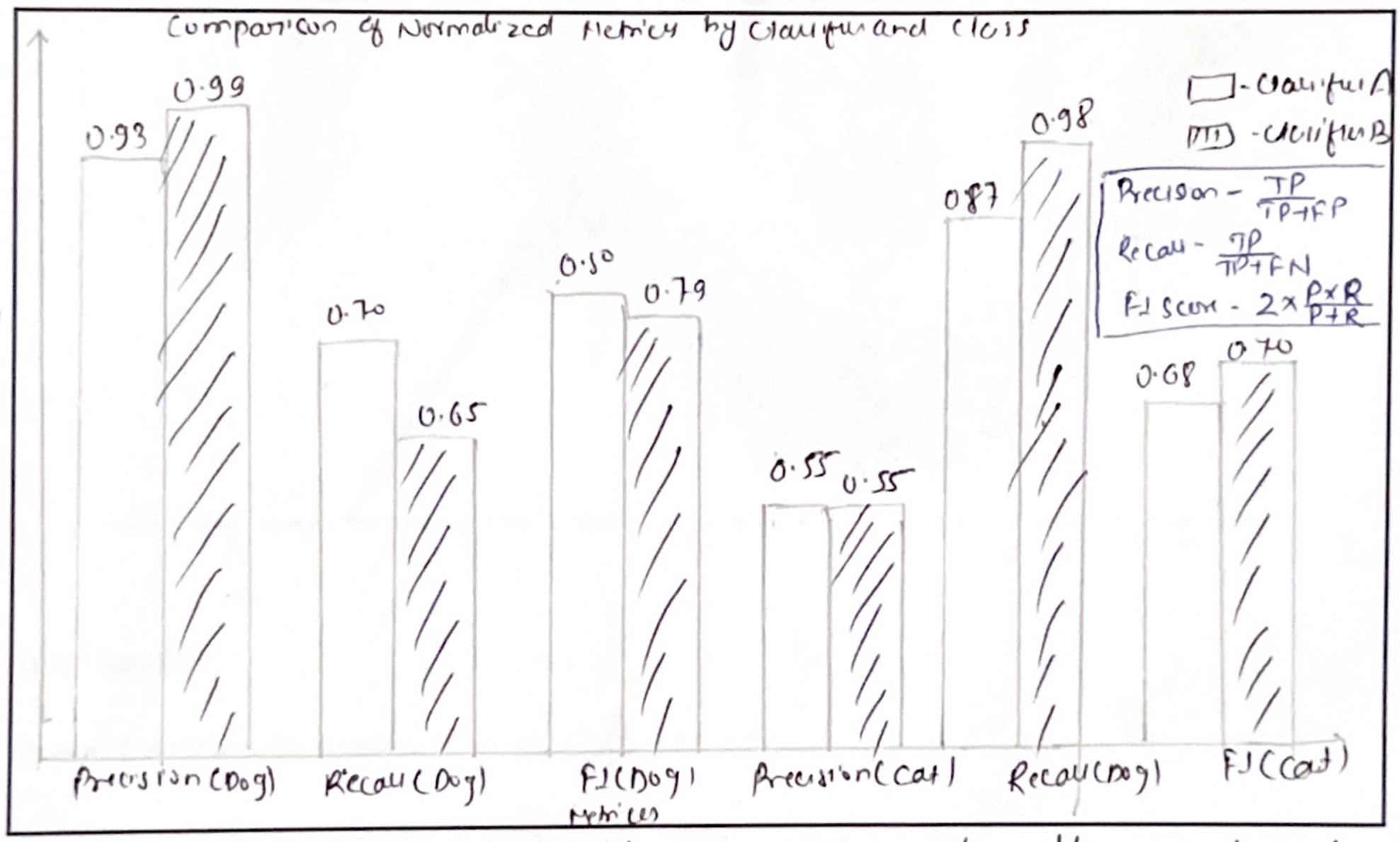
Task 2C: Brainstorming New Ideas: [/ 3P]

Your thesis work entails generating innovative concepts to rectify the shortcomings you have detected. In this particular assignment, **present a Visualization modification you would like to add in order to enhance the proposed solution.** Alternatively, you can propose adjustments to the existing solution methodology to enable the comparison of outcomes from 20 distinct classifiers instead of limiting it to just 2.

Briefly describe the problem that you mean to solve:
(Indicate "2A" if you are presenting a solution for the issue outlined in task 2A.)

→ The bar chart in task 3b (discussed in task 2A) fails to properly account for the fact that there are more dog images than cat images, which could make the classifiers seem better at identifying dogs just because there are more of them. This could lead to wrong conclusions about how good the classifiers really are.

Proposed Modification:
(You can use words, a rough sketch, or both.)



→ To address this problem, we should replace raw counts with normalized metrics like precision, recall and F1 score.

Precision :- tells us how many of images identified as a certain class (eg dog) are actually correct.

Recall :- indicates how well the classifier is at identifying all instances of a class (eg. catching all the dog images)

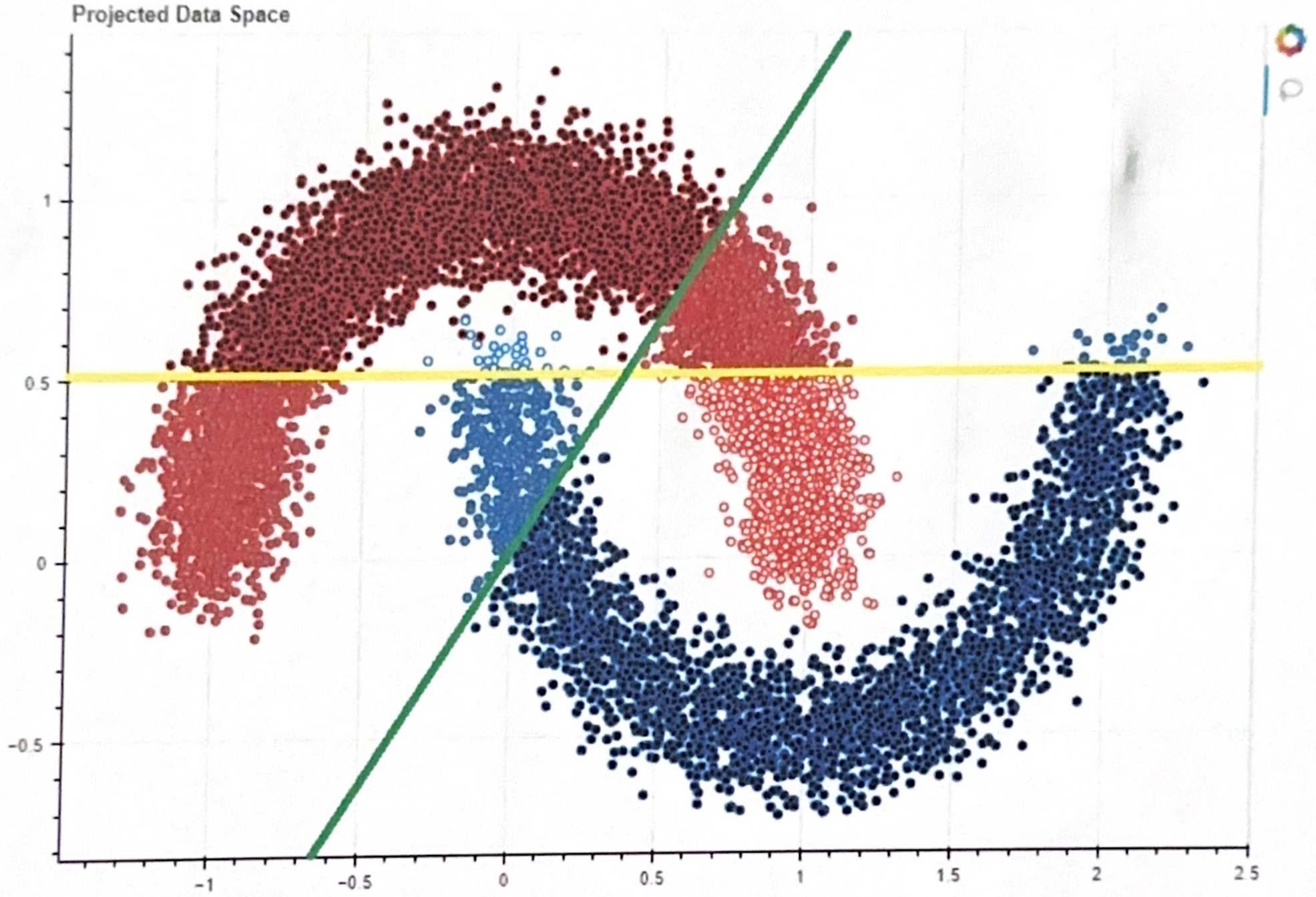
F1 score :- combines precision and recall into a single metric, balancing both concerns.

Using these metrics will give more accurate and fair comparison of how well the classifiers perform on both the dogs and cat classes, avoiding misleading interpretations caused by the larger number of dog samples.

We can even use (Decision Boundaries) to solve this issue by the help of visual clarity, identifying bias, and adjustment insights.

Task 2D: Visual Analysis - Gaining Insights: [/ 1P]

What purpose does a visualization serve if it doesn't allow for interpretation and insights? Both classifiers (A and B) employ a linear decision boundary represented by **green** and **yellow** lines in the scatterplot below to differentiate between cat and dog images. **Your task involves determining the association between the decision boundaries and their respective classifiers, based on your analysis of the given dataset.**



Your answer:

Green Decision Boundary: Classifier A (A/B)

Yellow Decision Boundary: Classifier B (A/B)