Thesis Eligibility Assessment

Name: **Anilkumar Bellamakondi**
Submission Deadline: **13.09.2024 23:59**
Requirement for Eligibility: Achieve a minimum score of **70%** (**14 points out of 20**)

Instructions:

- For the coding assignment (tasks 1A, 1B, and 1C):
  - Along with this PDF, you will also be provided with the dataset (in CSV format).
  - You are welcome to use JavaScript, Python, or any programming language you are familiar with. The visual representations presented within this assessment document are generated using the Bokeh visualization library in the Python programming language.
  - Please upload your code solution to a new GitHub repository.
  - Include screenshots of your solutions in the README.
  - Ensure that your GitHub repository is set to public.
- For the remaining tasks, kindly print out this PDF, complete your responses, and subsequently share a scanned PDF copy with us **(only Pages 6, 7, and 8 required)**, along with the **GitHub link**.
- Kindly refrain from sharing or discussing these questions/your solutions with others.

For grading only:

Task ID	Task	Points	Result
1A	Implementation - Data Manipulation	3	
1B	Implementation - Custom Visual Encoding	4	
1C	Implementation - Linked Interactivity	6	
2A	Critical Thinking Assessment	2	
2B	Literature Search	1	
2C	Brainstorming New Ideas	3	
2D	Visual Analysis - Gaining Insights	1	
Sum		20	

Remark: ______________________

Problem statement:
Imagine a dataset of 10,000 images comprising 7,000 dog images and 3,000 cat images. These images are projected onto a 2-dimensional plane. Subsequently, we train two distinct classifiers, namely Classifier A and Classifier B, on this projected dataset to distinguish between dog and cat images. Surprisingly, both classifiers yield an identical accuracy of 75%. However, a more detailed examination is desired. Our objective is:

1. To determine whether the data points (images) are accurately classified by both classifiers, accurately classified by only one of the classifiers, or not accurately classified by either of them.
2. To assess the performance of these two classifiers concerning the label classes ("dog" and "cat").
3. To conduct a more thorough analysis by restricting our assessment to selected data points within the projected 2D space.

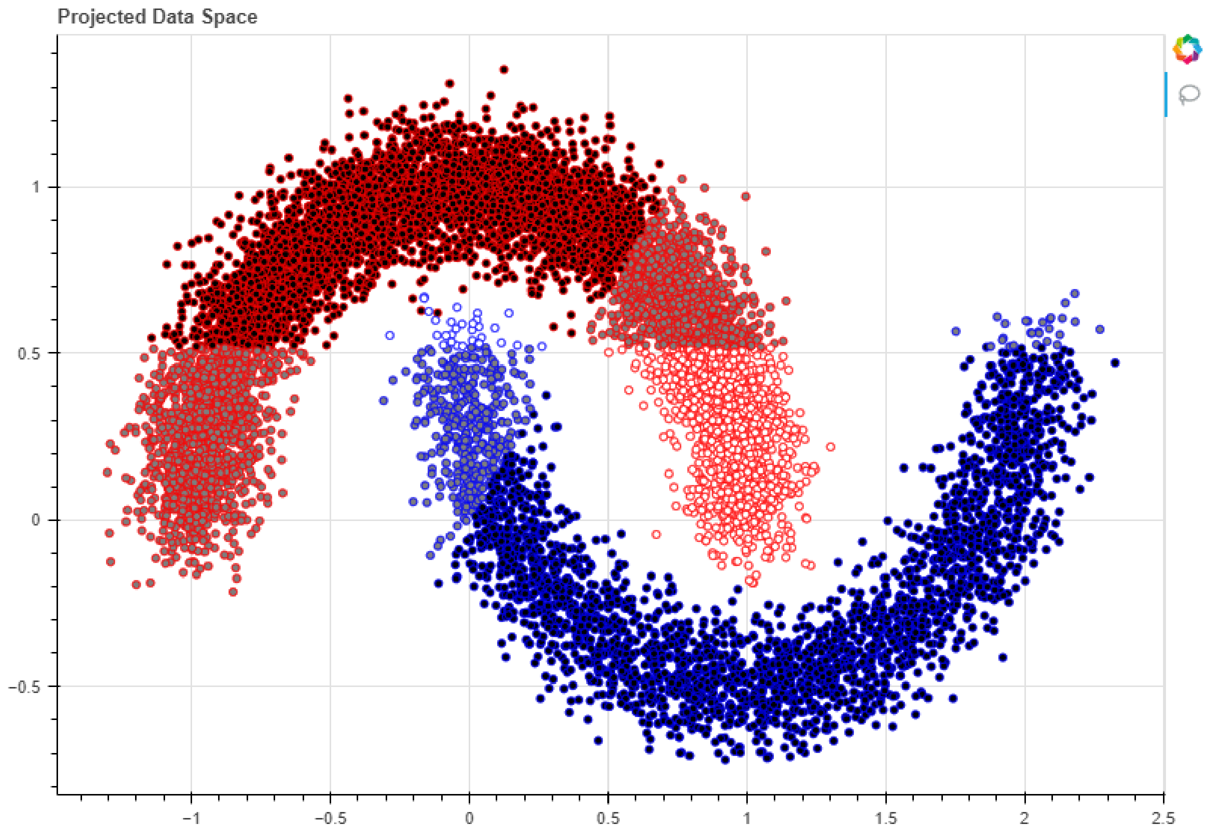
Details regarding the Dataset:

It is a synthetic dataset created for this assessment. It is in CSV format with 5 columns:

1. “**x**”: x-coordinate of the image in the projected data space.
2. “**y**”: y-coordinate of the image in the projected data space.
3. “**label**”: actual label of the image.
4. “**classifierA_predicted_label**”: label of the image predicted by classifier A.
5. “**classifierB_predicted_label**”: label of the image predicted by classifier B.

Proposed Solution:

For the 1st Objective, consider the following proposed solution:

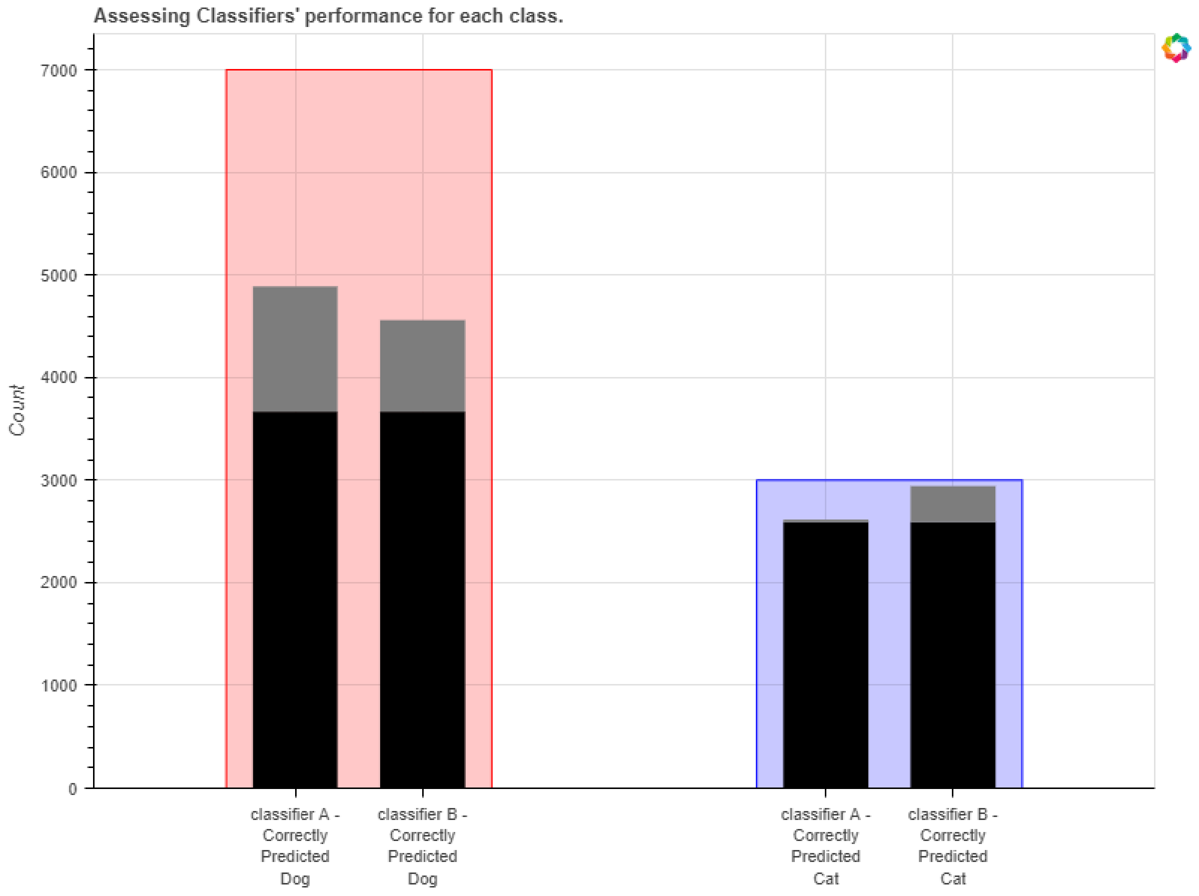


Circles(/Points) with **red borders** denote the images of dogs.
Circles(/Points) with **blue borders** denote the images of cats.
Black fill color denotes images(data points) accurately classified by both classifiers.
Gray fill color denotes images(data points) accurately classified by one of the classifiers.
White fill color denotes images(data points) accurately classified by none of the classifiers.

Task 1A: Implementation - Data Manipulation: [/ 3P]

Develop code that generates a scatterplot meeting these specifications. This involves skill in manipulating data to classify data points into three categories: those correctly classified by both classifiers, those correctly classified by only one of the classifiers, and those not accurately classified by either. **It's important to note that an identical solution isn't mandatory.** Instead, you have the freedom to devise a variation while ensuring it can differentiate between dog/cat labels for each data point. Additionally, it should indicate the specific category to which each data point belongs out of the three mentioned.

For the 2nd Objective, consider the following proposed solution:

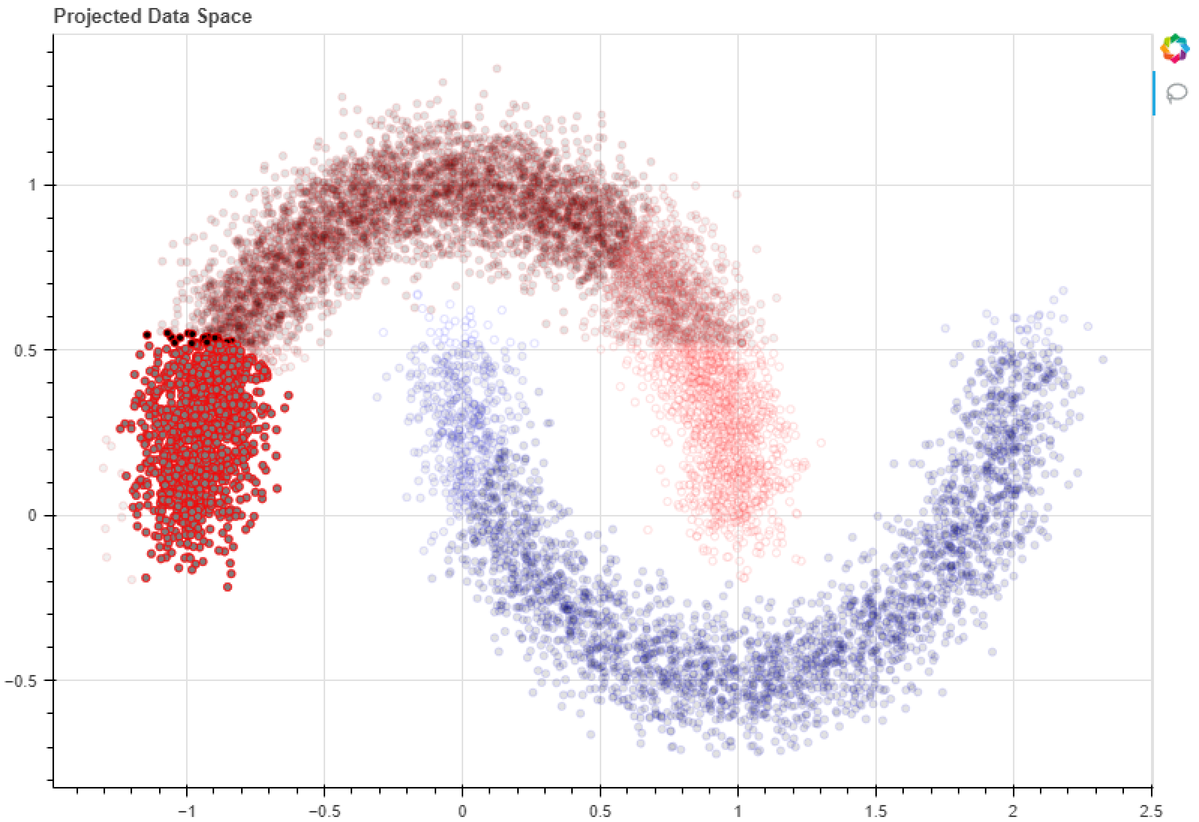


Colors have similar meanings as described in the previously shown scatterplot. The rest is self-explanatory.

Task 1B: Implementation - Custom Visual Encoding: [/ 4P]

Compose the code required to produce the customized bar chart as shown. A key strength of employing a programming language to craft visualizations (in contrast to using general visual analysis tools) lies in its capacity to generate unique visual representations. We aim to examine this capability in this context.

For the 3rd Objective, consider the following proposed solution:



The displayed scatterplot highlights a set of points that have been selected utilizing the Lasso Selection tool. What remains unillustrated here is how the act of selecting specific points leads to modifications in the customized bar chart.

Task 1C: Implementation - Linked Interactivity: [/ 6P]

At present, the gray-fill points, which correspond to images accurately classified by only one of the classifiers, do not indicate whether the correct classification came from classifier A or classifier B. To address this informational gap, we introduce linked interactivity between the two visual representations. When specific points on the scatterplot are chosen, the customized bar chart adjusts to display counts relevant to the selected data points instead of the entire dataset. In other words, **you need to link the two visualizations described in Task 1A and 1B via filter-based interaction.** In this task, your objective is to implement this type of linked interactivity, a common feature in many dashboards. It's important to note that you are free to employ any selection tool, such as the box selection tool, and aren't confined to using only the lasso selection tool.

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.

Overreliance on Accuracy:
Neglecting precision and Recall : Precision and recall are more nuanced metrics that consider false Positives and false negatives. A Visualization might not adequate highlight differences in these metrics leading to misleading Conclusions.

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. No points will be granted if the reference provided is generic.

claVis: An Interactive Visual Comparision System for Classifiers by (Frank Heyen, Tanja Munz and Colleagues).

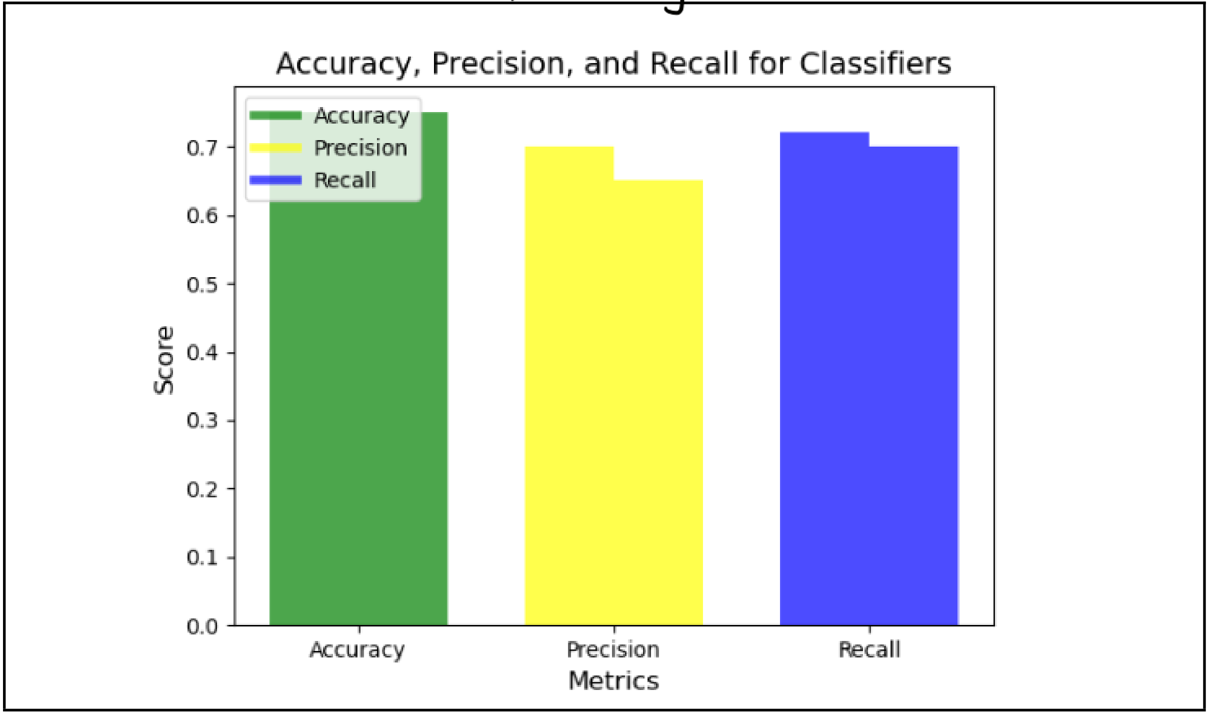
Task 2C: Brainstorming New Ideas: [/ 3P]

Your thesis work entails generating innovative concepts to rectify the shortcomings you have detected. In this particular assignment, **present a <u>Visualization modification</u> 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:
(Just indicate "2A" if you are presenting a solution for the issue outlined in task 2A.)

"2A" Visualization neglects precision and recall, which are crucial metrics for evaluating classifier performance, particularly in case of class imbalance. Over reliance on accuracy can lead to misleading conclusions, as it overlooks false positives and false negatives.

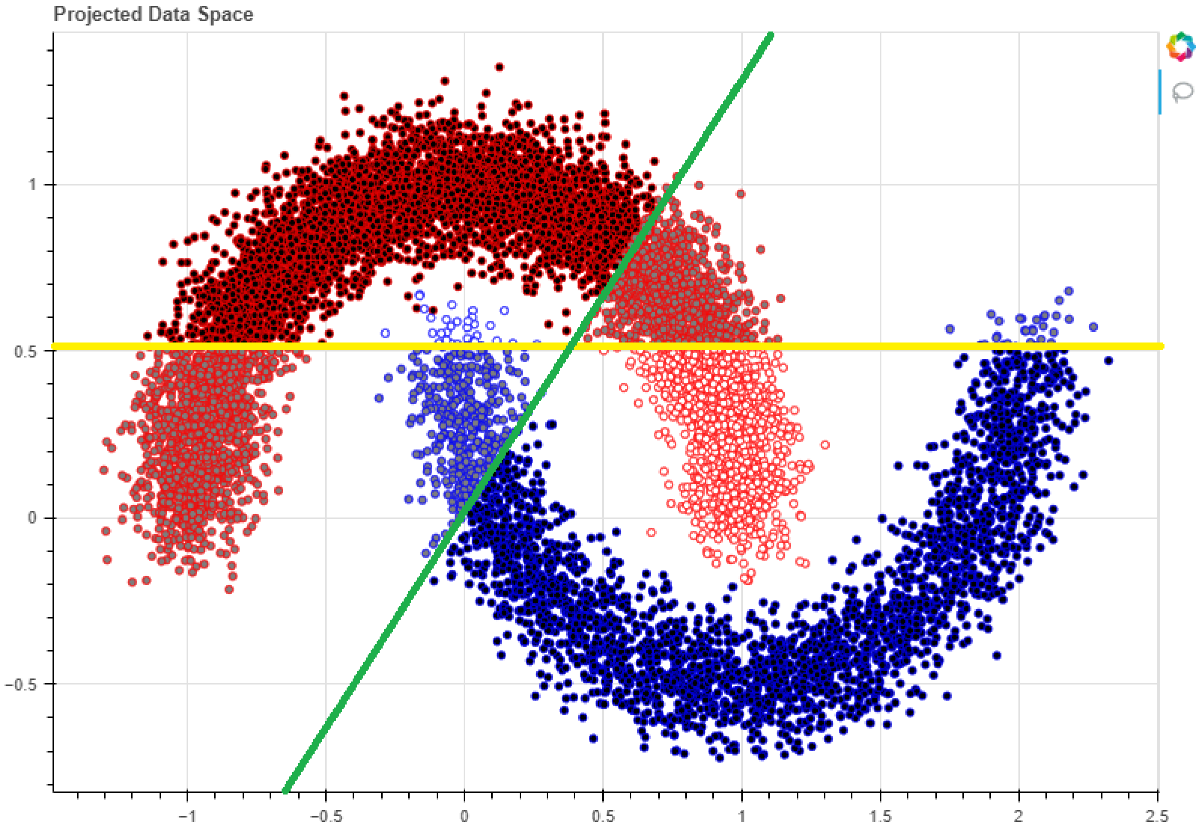
Proposed Modification:
(Sketch + brief description)



To address the issue, I propose adding Precision & recall metrics alongside accuracy in visualization. This is done using Bar chart with 3 metrics; accuracy, Precision & recall, displayed for each classifier. This modification would allow users to quickly identify discrepencies between the metrics, especially in senarios where one classifier may excel in accuracy but perform poorly in precision or recall.

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)