Interpretation

1 - In Figure 1 we have a data distribution, the dots represent the sparse data for the axis X and Y, and the lines represent the fit of a hypothetical classification model. Based on the distributions of Figure 1:

Which distribution has the best balance between bias and variance?

Describe your thoughts about your selection.

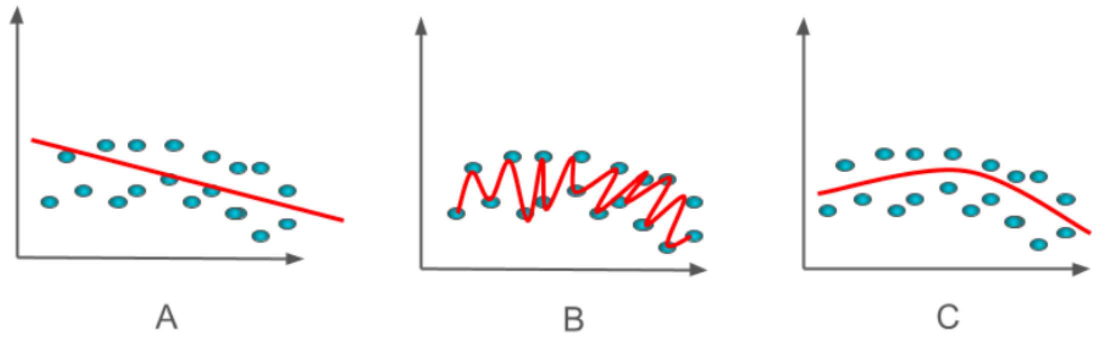
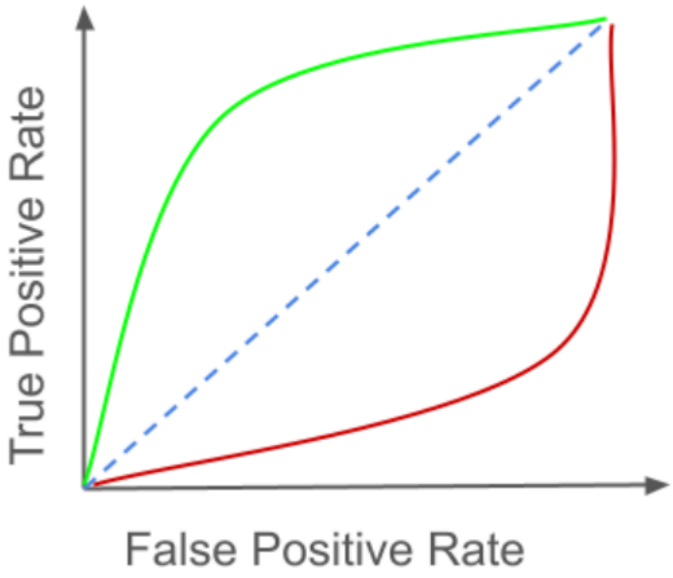


Figure 1 - Data distribution samples

Answer: Distribution C has the best balance.Bias refers to the error introduced by approximating a real-world problem (which is usually complex) with a simplified model. A high-bias model makes strong assumptions about the data and tends to underfit and about the Variance refers to the model's sensitivity to fluctuations in the training data. A high-variance model is very sensitive to the training data and tends to overfit, capturing noise rather than the underlying pattern.

- Distribution A will present a high bias,because the model its too simple
- Distribuition B Will present a High Variance cause the model is too Complex,overfitting the training data
- The curve in C follows the general trend of the data without being overly sensitive to individual data points. It captures the underlying relationship without being overly simplistic (like A) or overly complex (like B).

2 - Figure 2 presents a simple graph with 2 curves and 1 line. In model selection and



evaluation:

- What is the purpose of this graph and its name?

  Answer : The graph is called a Receiver Operating Characteristic (ROC) curve. Its purpose is to visualize the performance of a binary classification model at various classification thresholds. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR).

- What kind of model result does the dashed line represent?

Answer : The dashed line represents a random classifier or a model with no predictive power. It has an Area Under the Curve (AUC) of 0.5. Essentially, it's as good as flipping a coin.

Which curve represents a better fit, the red or the green? Why?

Answer : The green curve represents a better fit.The Red line goes to the bottom right corner, in that area, the model will classify more negative cases as positives (False Positives), for that reason, the green curve is the best, it classifies more positives cases correctly with few false positives, in comparison with the other.

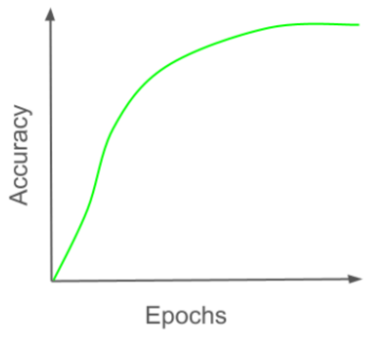
- Describe your thoughts about your selection.

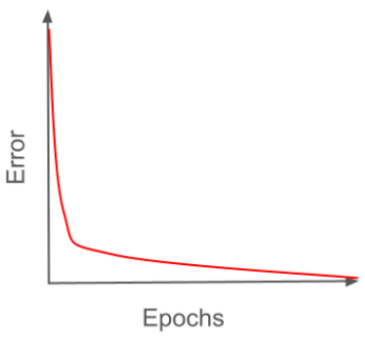
  Answer : A good model aims to maximize the True Positive Rate (correctly identifying positives) while minimizing the False Positive Rate (incorrectly identifying negatives as positives).

  The overall performance of a classifier is often summarized by the Area Under the Curve (AUC). A perfect classifier has an AUC of 1.0 (it would hug the top-left corner of the graph).

The green curve is closer to the top-left corner, indicating a higher True Positive Rate for a given False Positive Rate across various thresholds. It has a larger AUC than the red curve, signifying better overall performance. The red line, it will perform worse than random chance.

3 - Figure 3 presents a classification model training and the evaluation. This model classifies

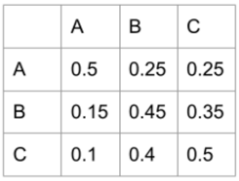




3 classes (aph B

represents on of the

model usin es trained.



	A	B	C
A	0.5	0.25	0.25
B	0.15	0.45	0.35
C	0.1	0.4	0.5

- Can we say that the model has a good performance in the test evaluation?

Answer : No, the model does not have good performance in the test evaluation.

- What phenomenon happened during the test evaluation?

Answer : The model presents an unbalanced dataset.

- Describe your thoughts about your selection.

Answer : The accuracy increases rapidly and plateaus, suggesting the model quickly learns the training data. This could be a sign of overfitting, especially if validation accuracy were not also tracked.

The error decreases rapidly and approaches zero, further reinforcing that the model is learning the training data very well. Again, this could indicate overfitting if the validation loss doesn't follow the same trend.

While the model appears to perform well on the training data (high accuracy, low loss), the confusion matrix reveals that the class performance is relatively poor in all of the 3 class, at best 50% of acurracy. This indicates that the model hasn't generalized well to unseen data, even though it appears good on the training set, the dataset is unbalenced.