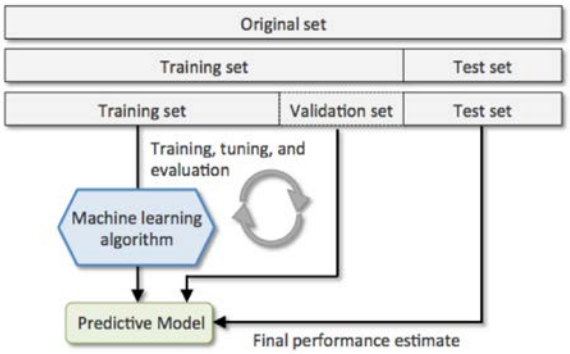
Questions to critical GxP AI/ML applications

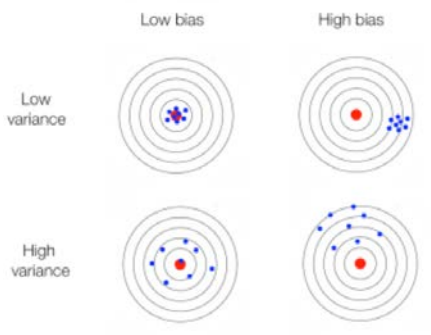
Based on static AI/ML algorithms and supervised learning

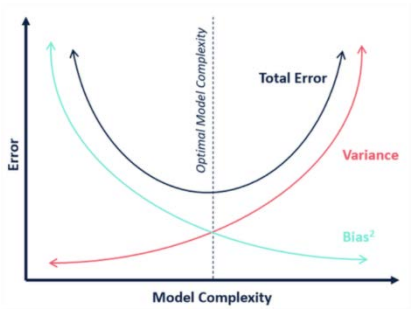
Datasets



1. What was the process for training, validation (optimizing) and testing of the algorithm, who was involved in the different phases and which deviations were made from the plan, if any?
2. How large are the datasets used to train, validate and test the algorithm, how is each individual data element within each group identified (named) and where are the datasets stored?
3. What evidence exists that no part of the dataset used to test the algorithm has previously been used to train or validate (optimize) the same algorithm, or originates from the same subjects?
4. When were the data points used to test the algorithm separated from the complete pool of data points and what selection criteria were used?
5. What kind of data cleaning, normalization, homogenization, exclusion criteria, data synthetization or similar were the test data subject to and why?
6. How was it ensured that the test dataset is representative of real data from the intended scope of the application and contains enough challenging data (e.g. Siberian Husky or wolf)?
7. What features in the training dataset have the highest effect on the output of the algorithm, how has that influenced the selection of, and how does it correspond to the test dataset?
8. How was it ensured that the test dataset covers any technical differences (e.g. formatting) which may arise in real data due to differences in personnel, processes and equipment?
9. How was the correct classification verified of the data used to train, validate and especially test the algorithm, and has the classification e.g. been verified by a second person or by laboratory tests?
10. How old is the test data and is it still relevant? Is the algorithm's F1 Score likely to deteriorate due to changes in input data over time (e.g. changes health data due to closed fitness centres during COVID-19 lock-down), and if so, what are the plans for retraining and calibration of datasets?

Bias and Variance





11. How was the algorithm optimized to deal with bias and variance (bias – variance tradeoff), what is the result of this optimization as seen in the test and how do bias – variance tradeoff graphs look?
    a. Bias is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
    b. Variance is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).
12. How is system performance monitored across consecutive training, validation and test cycles? E.g. F1 Score going from 70 - 80 - 90% may indicate an over confident algorithm due to overfitting.

Confusion Matrix and Metrics

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

13. What are the values in the confusion matrix (TP/FP/FN/TN) and the following metrics?
    a. Sensitivity [TP/(TP+FN)] is the number of items correctly identified as positive out of total true positives. Sensitivity is also called Recall, Hit Rate or True Positive Rate (TPR)
    b. Specificity [TN/(TN+FP)] is the number of items correctly identified as negative out of total true negatives. Specificity is also called Selectivity or True Negative Rate (TNR)
    c. Precision [TP/(TP+FP)] is the number of items correctly identified as positive out of the total identified as positive. Precision is also called Positive Predictive Value (PPV).
    d. False Positive Rate [FP/(TN+FP)] is the number of items wrongly identified as positive out of total true negatives. E.g. a man being declared as pregnant. Is also called Type I error.
    e. False Negative Rate [FN/(TP+FN)] is the number of items wrongly identified as negative out of total true positives. E.g. pregnant woman declared as not pregnant. Called Type II error.
    f. Accuracy [(TP+TN)/(N+P)] is the percentage of total items classified correctly. Should not be used with uneven sets of classes, as accuracy of one class can overpower the other.
    g. F1 Score [2*(Precision*Sensitivity)/(Precision + Sensitivity)] is a harmonic mean of Precision and Sensitivity. The F1 Score has the advantage over accuracy that with uneven classes it gives a better metric to calculate the model performance.

Interpretation of Results

14. Which of the quadrants of the confusion matrix and which of the metrics are more important for the intended scope of the application and if low scores are seen, why may these be less important?
15. How was the intended scope of the application defined and limited based on both test data and test results including the confusion matrix and metrics?
16. How was the threshold defined for end results, e.g. is an outcome of ’50.01% it is a dog’ interpreted to the result ‘it is a dog’, and when, if ever, would an outcome require human interaction?