Problem statement

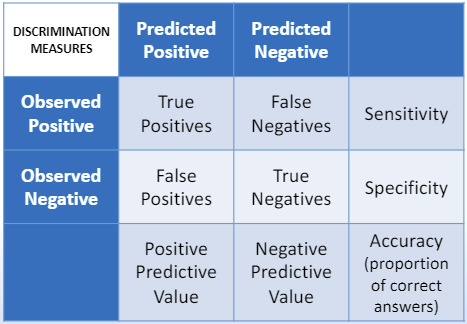
We have an algorithm to classify fetal health into three classes:1.Normal 2.Suspect 3.Pathologicalusing features extracted from Cardiotocogram exams.The performance of this algorithm in an external validation dataset is contained in file:pred\_fetal\_health.csv

The file contains 1701 rows (1 per patient) and 2 columns:y\_true (equal to 0 for Normal fetal health and 1 for Suspect or Pathological status)p\_pred (refers to the probability of Suspect or Pathological status as estimated by the predictive algorithm)

Based on this information, you are asked to provide information on the predictive performance of the algorithm. Note: In order to answer the questions in this tutorial quiz you can use any programming language of your choice such as excel. I recommend that you use a python code similar to the one we used during the tutorial session.

Question 1

Assuming a standard probability threshold of 0.5, classify fetal health into healthy or otherwise (pathological or suspect) and fill in the discrimination measures table below



|  |  |  |  |
| --- | --- | --- | --- |
| **Discrimination Measures** | **Predicated Positive** | **Predicated Negative** |  |
| **Observed Positive** | True Positives (TP) = 230 | False Negatives (FN) = 61 | Sensitivity = 0.79 |
| **Observed Negative** | False Positives (FP) = 98 | True Negatives (TN) = 1312 | Specificity = 0.93 |
|  | Positive Predictive Value (PPV) = 0.70 | Negative Predictive Value (NPV) = 0.96 | Accuracy = 0.91 |

Question 2

Draw the ROC (TPR against FPR) and provide the AUC

A graph of a function

Description automatically generated with medium confidence

Question 3

You want your algorithm to operate at a point where sensitivity is high, around 90% (SEN=0.9) since you do not want to have false negatives (failing to identify problems with fetal health).

Find the threshold at which sensitivity is closest to 0.9, and repeat the exercise in Question 1 for this threshold.

Tip: Think of minimising the distance: x = SEN-0.9 (similarly to what we did in the tutorial when finding the threshold that maximised g-mean)

Best Threshold for Sensitivity ≈ 0.9: 0.28； gmean=:0.87, F1: 0.69

|  |  |  |  |
| --- | --- | --- | --- |
| **Discrimination Measures** | **Predicated Positive** | **Predicated Negative** |  |
| **Observed Positive** | True Positives (TP) = 262 | False Negatives (FN) = 29 | Sensitivity = 0.90 |
| **Observed Negative** | False Positives (FP) = 209 | True Negatives (TN) = 1201 | Specificity = 0.85 |
|  | Positive Predictive Value (PPV) = 0.56 | Negative Predictive Value (NPV) = 0.98 | Accuracy = 0.86 |

Question 4

Explain the difference between the model discrimination if question 1 and the model discrimination in question 3

Model in question 1 represents a balanced approach, prioritizing both sensitivity and specificity with a threshold of 0.5. It performs with high specificity and accuracy but lower sensitivity.

Model in Question 3 focuses on maximizing sensitivity (around 90%), which decreases the number of false negative but at the cost of higher false positive. This adjustment is preferred in the scenario where missing a true positive (false negative) is more critical than having some false positives.

The choice of threshold is essential in clinical decision-making, it should depends on the specific application and the relative importance of avoiding false negatives versus false positives. This is usually a kind of trade-off.

Question 5

Plot a calibration curve using 5 bins and calculate the Brier score.

A graph with a line

Description automatically generated

Brier Score: 0.0706

Question 6

Briefly comment on the calibration of the model probabilities. Does the model provide a good prediction for patients across all levels of risk? Does the model overestimate or underestimate patients’ risks?

The dashed line represents perfect calibration, where predicted probabilities perfectly match the actual outcomes. In general, the calibration curve is close to the dashed line, it indicates that the model's predictions are well-calibrated. But the calibration curve is always below the dashed line, especially between 0.4 to 0.6, indicating the model is **overestimating** the risk. The actual observed frequency of the event is lower than the predicted probability.

The calibration curve shows that the model is well-calibrated for most parts of the probability range, especially at lower and higher probabilities.

The Brier Score is a measure of how close the predicted probabilities are to the actual outcomes. It ranges from 0 to 1, with lower values indicating better accuracy and calibration. The Brier Score of 0.0706 suggests that the model's probabilistic predictions are generally accurate overall.

This analysis shows that while the model performs well, there is room for improvement in specific probability ranges (0.4 - 0.6) to achieve even better calibration and predictive performance.