AutoML Modeling Report

Jessica Ofoh

Binary Classifier with Clean/Balanced Data

Train/Test Split How much data was used for training? How much data was used for testing?	In this first case, the number of total images used were 200 (100 for pneumonia and normal), where 152 were utilized for training, 22 images for testing and 26 images for validation by the Google AutoML platform.
Confusion Matrix What do each of the cells in the confusion matrix describe? What values did you observe (include a screenshot)? What is the true positive rate for the “pneumonia” class? What is the false positive rate for the “normal” class?	Confusion matrix refers to a tabular layout that visualizes the performance of a classification algorithm by mentioning the hit values for every scenario [true positive (TP), false negative (FN), true negative (TN) and false positive (FP)]. Moreover, here, the values observed are 89.3% and 10.7% for TP and FP respectively. 
Precision and Recall What does precision measure? What does recall measure? What precision and recall did the model achieve (report the values for a score threshold of 0.5)?	The value of precision reflects the amount (fraction) of positive predictions that were accurately positive from all positive predictions; while recall refers to the fraction of positive predictions that were accurately positive from all real positive labels (where here positive refers to pneumonia cases). Moreover, at a score threshold of 0.5, 93.2% turned out to be the value for both precision and recall.

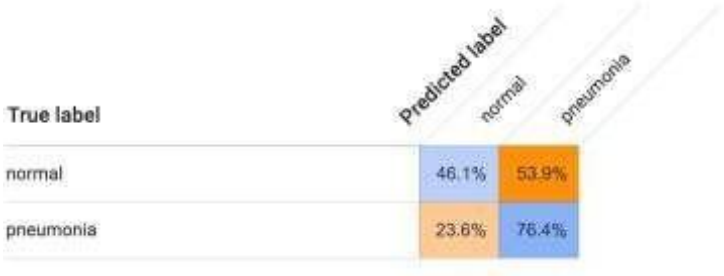
Score Threshold When you increase the threshold what happens to precision? What happens to recall? Why?	Upon an increase in the value of threshold (say 0.5 to 0.9), the value of precision tends to increase in general. However, in the same scenario, the value of recall tends to decrease. This occurs because when the value of threshold is low, the model tries to classify more images which might lead to more misclassified labels. This isn't the case when the value of threshold is high, as in that case, the model tries to classify fewer images to be more confident in prediction that leads to fewer misclassifications.

Binary Classifier with Clean/Unbalanced Data

Train/Test Split How much data was used for training? How much data was used for testing?	In this first case, the number of total images used were 400 (300 for pneumonia and 100 for normal), where 305 were utilized for training, 52 images for testing and 43 images for validation by the Google AutoML platform.
Confusion Matrix How has the confusion matrix been affected by the unbalanced data? Include a screenshot of the new confusion matrix.	After making the change in the dataset (adding 200 more images of pneumonia), the confusion matrix changed significantly. Intricately, the value of TP increased (93.6%) which meant that there were more correct positive identifications. Moreover, the value of TN decreased (94.3%), which meant that the tendency to correct identification of normal images decreased. 

Precision and Recall How have the model's precision and recall been affected by the unbalanced data (report the values for a score threshold of 0.5)?	Yes, the value of both precision and recall changed after making the changes in the dataset for this case. At 0.5 threshold value, the precision and recall both turned out to be 94.7% which is an increase from the previous case.
Unbalanced Classes From what you have observed, how do unbalanced classed affect a machine learning model?	As perceived from the results, it's clear that an unbalance classed dataset brings BIAS in the machine learning model. This refers to the fact that an ML model has a tendency to classify more input objects (here, images) into classes that are greater in number in the training process. This is clear as here, the model classifies more images as 'pneumonia' because of greater number of training images for the same.

Binary Classifier with Dirty/Balanced Data

Confusion Matrix How has the confusion matrix been affected by the dirty data? Include a screenshot of the new confusion matrix.	After making the change in the dataset (providing balanced, yet mislabeled 'dirty' data to ML model), the confusion matrix changed significantly in a negative way. Intricately, the performance of the ML model was poor in every aspect of the prediction making process which can be clearly seen in the snapshot below. This was in regards to mislabeled and wrong data fed to the model, which then led to ML model making wrong decisions. 

Precision and Recall How have the model's precision and recall been affected by the dirty data (report the values for a score threshold of 0.5)? Of the binary classifiers, which has the highest precision? Which has the highest recall?	Again, the value of both precision and recall changed after making changes in the dataset for this case; where at 0.5 threshold value, the precision and recall both turned out to be 64.9% which is significantly bad performance. Moreover, based upon comparative deduction of performance, the 'ML Model trained from clean and unbalanced data' provided the highest value of precision and recall.
Dirty Data From what you have observed, how does dirty data affect a machine learning model?	The introduction of mislabeled (dirty) data brings a lot of contradictions for ML models to deal with after training. This results in several misclassifications by the model as the data is wrong in nature, which results in its poor performance.

3-Class Model

Confusion Matrix Summarize the 3-class confusion matrix. Which classes is the model most likely to confuse? Which class(es) is the model most likely to get right? Why might you do to try to remedy the model's "confusion"? Include a screenshot of the new confusion matrix.	The 3-class confusion matrix here portrays the classification results of all three classes (like the 2 × 2 matrix) to understand ML model's performance. Moreover, as deducted from the matrix, the model is likely to confuse in the case of 'normal' class and 'virus' class for prediction, while the model is likely to get 'bacteria' class right almost every single time. Moreover, we might need to remedy the confusion here (to improve performance) by increase the number of images in the training dataset for each class in order for ML models to understand the patterns in images for accurate predictions. 

Precision and Recall What are the model's precision and recall? How are these values calculated (report the values for a score threshold of 0.5)?	In the final scenario with 3-class confusion matrix, the precision and recall values turn out to be 91.83% and 89.36%, at a 0.5 threshold score. Moreover, in the scenario of 3 × 3, in order to calculate the value of precision and recall, their individual values are calculated initially (for each class), which is followed by taking the average of those values. By doing so, the overall values of precision and recall can be determined.
F1 Score What is this model's F1 score?	The F1 score of this model turns out to be 0.906 (0.905781644), and is a score which can be calculated with the following equation: $$F1 = 2 \times \frac{\overline{P \times R}}{P + R}$$ where P and R refer to precision and recall respectively.