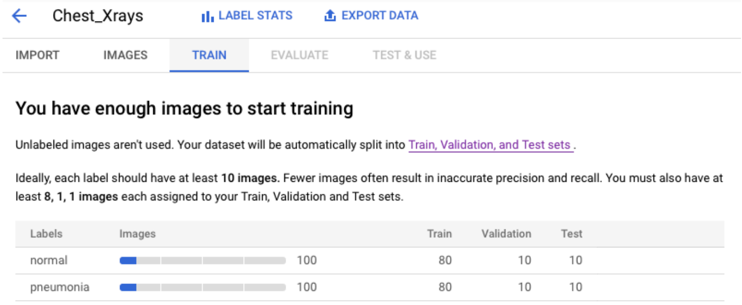
AutoML Modeling Report

Italo Farfán

Binary Classifier with Clean/Balanced Data

Train/Test Split
How much data was used for training? How much data was used for testing?

80% for train and 20 % testing (10% validation and 10% test) as show in the figure below.



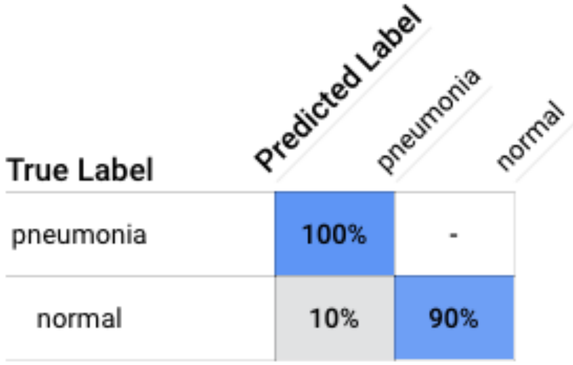
You have enough images to start training

Unlabeled images aren't used. Your dataset will be automatically split into Train, Validation, and Test sets .

Ideally, each label should have at least **10 images.** Fewer images often result in inaccurate precision and recall. You must also have at least **8, 1, 1 images** each assigned to your Train, Validation and Test sets.

Labels	Images	Train	Validation	Test
normal	100	80	10	10
pneumonia	100	80	10	10

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?

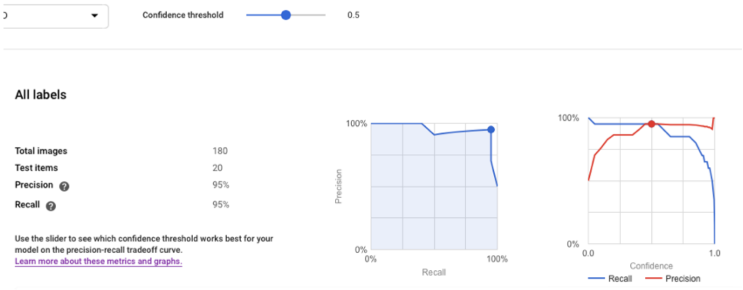
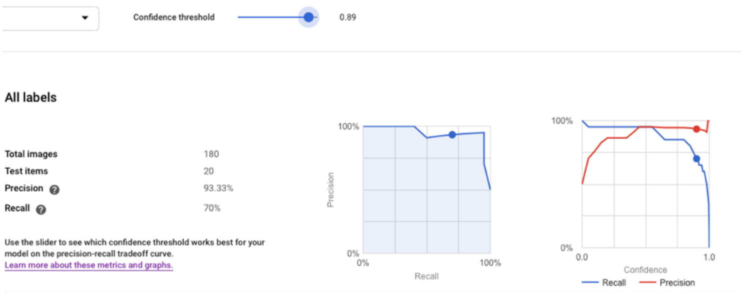


True Label	Predicted Label pneumonia	normal
pneumonia	100%	-
normal	10%	90%

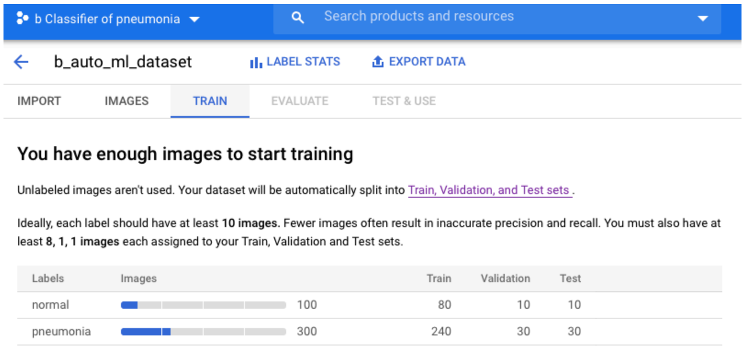
The Confusion Matrix represents the outcome of model evaluation. We can see the corrects and incorrect predictions.
The True Positive Rate = 100%
The False Positive Rate for the "normal"= 10%

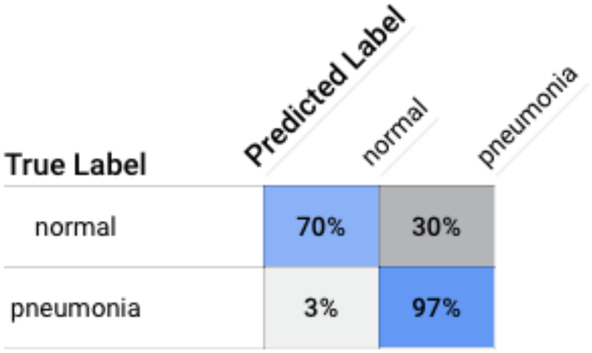
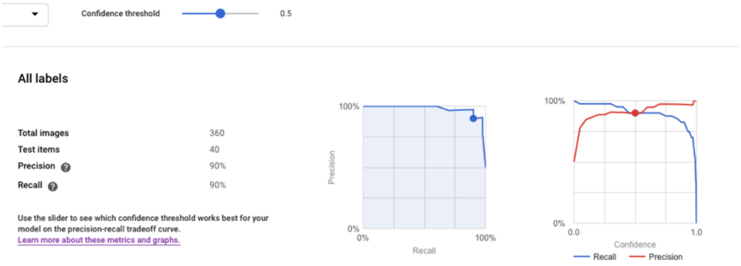
Precision and Recall
What does precision measure?

Precision measure the ratio of true positives to predicted positives.

What does recall measure? What precision and recall did the model achieve (report the values for a score threshold of 0.5)?	Recall measure the ratio of true positives to actual positives. Results of my model with Threshold of 0.5: Precision = 95% Recall = 95 % For other hand I like so much the definition from Wikipedia: Recall in this context is also referred to as the true positive rate or sensitivity, and precision is also referred to as positive predictive value (PPV); other related measures used in classification include true negative rate and accuracy. True negative rate is also called specificity.  Confidence threshold 0.5 All labels Total images 180 Test items 20 Precision 95% Recall 95% Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve. Learn more about these metrics and graphs. 100% Precision 0% 0% 100% Recall 100% 0% 0.0 1.0 Confidence Recall Precision Note: A high precision model produces fewer false positives. A high recall model produces fewer false negatives
Score Threshold When you increase the threshold what happens to precision? What happens to recall? Why?	Precision goes down and Recall goes down when I increase the threshold to 0.89 but if I continue increase the Threshold Precision goes up and Recall goes down as show in the figure below.  Confidence threshold 0.89 All labels Total images 180 Test items 20 Precision 93.33% Recall 70% Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve. Learn more about these metrics and graphs. 100% Precision 0% 0% 100% Recall 100% 0% 0.0 1.0 Confidence Recall Precision

Binary Classifier with Clean/Unbalanced Data

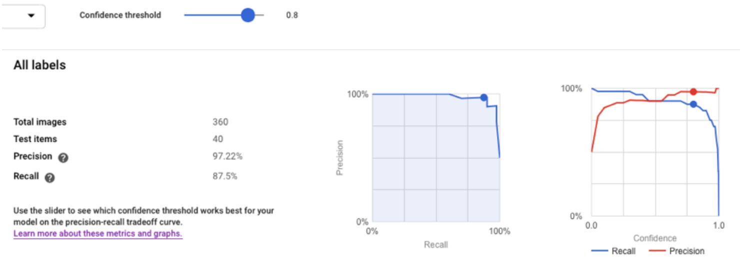
<table>
<tr>
<td><b>Train/Test Split</b>
How much data was used for training? How much data was used for testing?</td>
<td>80% for train and 20 % testing (10% validation and 10% test) as show in the figure below.


<b>You have enough images to start training</b>
Unlabeled images aren't used. Your dataset will be automatically split into Train, Validation, and Test sets .
Ideally, each label should have at least <b>10 images.</b> Fewer images often result in inaccurate precision and recall. You must also have at least <b>8, 1, 1 images</b> each assigned to your Train, Validation and Test sets.
<table>
<tr><th>Labels</th><th>Images</th><th>Train</th><th>Validation</th><th>Test</th></tr>
<tr><td>normal</td><td>100</td><td>80</td><td>10</td><td>10</td></tr>
<tr><td>pneumonia</td><td>300</td><td>240</td><td>30</td><td>30</td></tr>
</table>
</td>
</tr>
<tr>
<td><b>Confusion Matrix</b>
How has the confusion matrix been affected by the unbalanced data? Include a screenshot of the new confusion matrix.</td>
<td>I see two important changes: there is an increase of False Positives and there is an increase of True Positives.

<table>
<tr><th>True Label</th><th>Predicted Label
normal</th><th>pneumonia</th></tr>
<tr><td>normal</td><td>70%</td><td>30%</td></tr>
<tr><td>pneumonia</td><td>3%</td><td>97%</td></tr>
</table>
</td>
</tr>
<tr>
<td><b>Precision and Recall</b>
How have the model's precision and recall been affected by the unbalanced data (report the values for a score threshold of 0.5)?</td>
<td>Precision and Recall decreased 5% with Threshold of 0.5.
Precision = 90%
Recall = 90%

Confidence threshold 0.5
<b>All labels</b>
<table>
<tr><td>Total images</td><td>360</td></tr>
<tr><td>Test items</td><td>40</td></tr>
<tr><td>Precision</td><td>90%</td></tr>
<tr><td>Recall</td><td>90%</td></tr>
</table>
Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve.
Learn more about these metrics and graphs.

</td>
</tr>
<tr>
<td><b>Unbalanced Classes</b>
From what you have observed, how do unbalanced classed affect a machine learning model?</td>
<td>The model improved the predictions of pneumonia cases This is easy to see in the matrix and it is obvious because now we have more samples of pneumonia cases but other hand unblanced classed increase the falses positives and</td>
</tr>
</table>

this is very bad. If we were desing a model to predict cancer and help to doctor to take a decission to operate or not, have an increase of Falses Positives is so bad.

We may use a Threshold of 0.8 to help our model but the correct way is use a balanced classed.



Confidence threshold 0.8

All labels

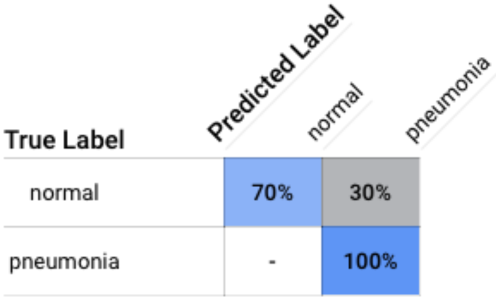
Total images	360
Test items	40
Precision	97.22%
Recall	87.5%

Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve.
Learn more about these metrics and graphs.

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.

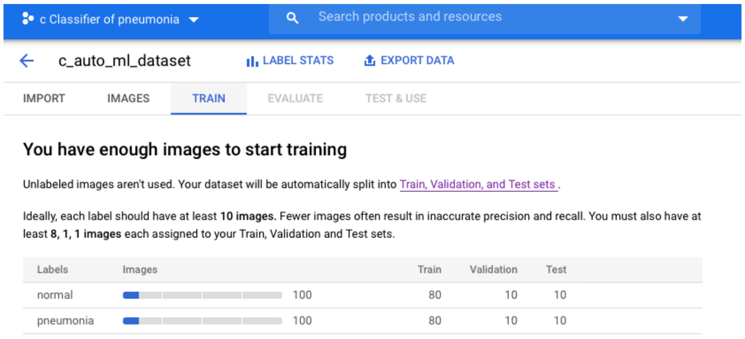
I see two important changes: there is an increase of False Positives and there is a decrease of True negative. This is very interesting because in my first model the False Negatives are 0%.



True Label	normal	pneumonia
normal	70%	30%
pneumonia	-	100%

Predicted Label

I used 80% for train and 20 % testing (10% validation and 10% test) as show in the figure below.

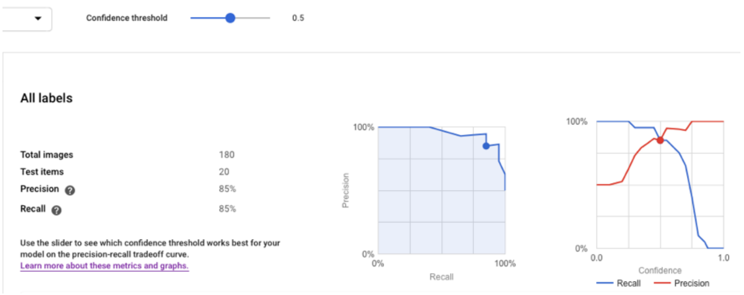
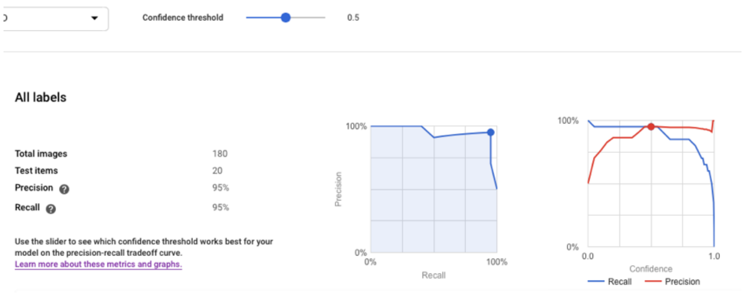
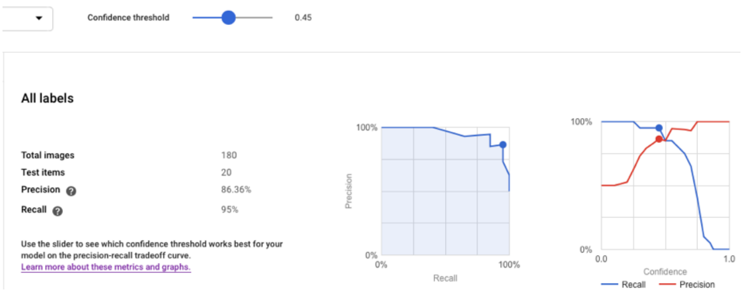


You have enough images to start training

Unlabeled images aren't used. Your dataset will be automatically split into Train, Validation, and Test sets .

Ideally, each label should have at least **10 images.** Fewer images often result in inaccurate precision and recall. You must also have at least **8, 1, 1 images** each assigned to your Train, Validation and Test sets.

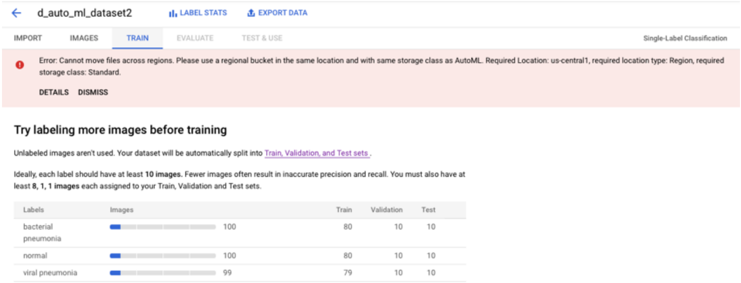
Labels	Images	Train	Validation	Test
normal	100	80	10	10
pneumonia	100	80	10	10

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?	The results of my model with Threshold of 0.5: Precision = 85% Recall = 85 %  Confidence threshold 0.5 All labels Total images 180 Test items 20 Precision 85% Recall 85% Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve. Learn more about these metrics and graphs. Precision Recall 100% 0% Confidence 0.0 1.0 Recall Precision Binary Classifier with Dirty/Balanced Data Of all binary classifiers the first model, The Binary Classifier with Clean/Balanced Data with 0.5 of Threshold, it has the highest precision and highest recall both 95%  Confidence threshold 0.5 All labels Total images 180 Test items 20 Precision 95% Recall 95% Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve. Learn more about these metrics and graphs. Precision Recall 100% 0% Confidence 0.0 1.0 Recall Precision Binary Classifier with Clean/Balanced Data
Dirty Data From what you have observed, how does dirty data affect a machine learning model?	In general, the performance is worse and I can see and understand the importance of good labeled process. We may use a Threshold of 0.45 to increase the recall but the precision still is low.  Confidence threshold 0.45 All labels Total images 180 Test items 20 Precision 86.36% Recall 95% Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve. Learn more about these metrics and graphs. Precision Recall 100% 0% Confidence 0.0 1.0 Recall Precision

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.

First it all I used 299 images. Something was wrong when I was uploading the 300 images and I had the same trouble many times, one of the images of viral pneumonia didn't upload correctly.



Error: Cannot move files across regions. Please use a regional bucket in the same location and with same storage class as AutoML. Required Location: us-central1, required location type: Region, required storage class: Standard.

Try labeling more images before training

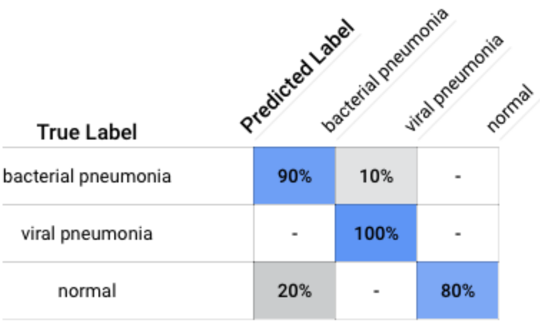
Unlabeled images aren't used. Your dataset will be automatically split into Train, Validation, and Test sets .

Ideally, each label should have at least **10 images**. Fewer images often result in inaccurate precision and recall. You must also have at least **8, 1, 1 images** each assigned to your Train, Validation and Test sets.

Labels	Images	Train	Validation	Test
bacterial pneumonia	100	80	10	10
normal	100	80	10	10
viral pneumonia	99	79	10	10

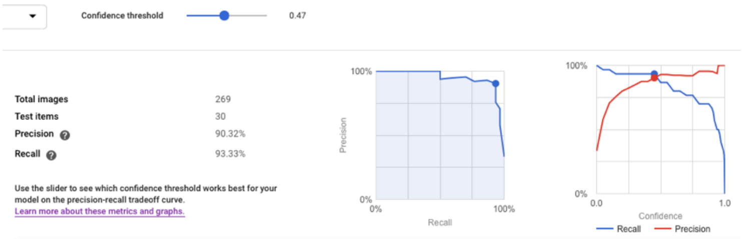
This is a good matrix because only two labels are bad predicted, but the first binary model still is better.

- The model most likely to confuse is the class "normal"
- The model most likely to get right is the class "viral pneumonia"



True Label	Predicted Label bacterial pneumonia	viral pneumonia	normal
bacterial pneumonia	90%	10%	-
viral pneumonia	-	100%	-
normal	20%	-	80%

To try to remedy the model's "confusion" we may use a Threshold of 0.47. But I would like to meet with a specialist to analyze images of viral pneumonia and images of healthy children to try to find some characteristic that will help us to better label the images.



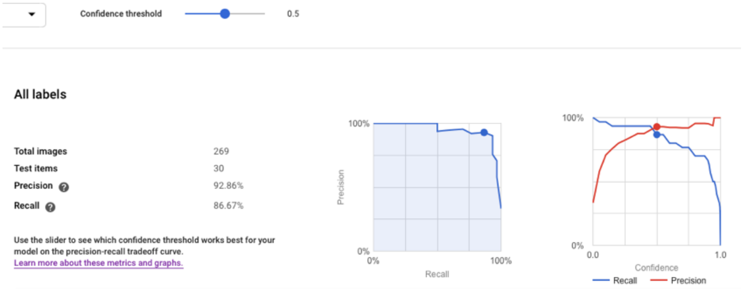
Confidence threshold 0.47

Total images	269
Test items	30
Precision	90.32%
Recall	93.33%

Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve.
Learn more about these metrics and graphs.

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)?

The results of my model with Threshold of 0.5:
Precision = 92.86%
Recall = 86.67 %



Confidence threshold 0.5

All labels

Total images	269
Test items	30
Precision	92.86%
Recall	86.67%

Use the slider to see which confidence threshold works best for your model on the precision-recall tradeoff curve.
Learn more about these metrics and graphs.

These values are calculated automatically by Google Cloud Platform, but we can calculate how we are learned in class:

- **Precision:** Ratio of true positives to predicted positives.

$$Precision = \frac{TP}{(TP + FP)}$$

- **Recall:** Ratio of true positives to actual positives.

$$Recall = \frac{TP}{(TP + FN)}$$

F1 Score
What is this model's F1 score?

First, remember that the F1-score is a function of precision and recall. We already learned how to compute the per-class precision and recall.
Here is a summary of the precision and recall for our three classes:

Class	Precission	Recall	F1-Score
bacterial pneumonia	82.82%	90%	85.71%
viral pneumonia	90.91%	100%	95.24%
normal	100%	80%	88.89%

The **macro-average F1-score, or F1** for short, is computed as a simple arithmetic mean of our per-class F1-scores.
So F1 = 89.95%