SCIO

December 21, 2020

Data preparation

The raw data consists on two dataframes, SCIO and DXA. The former contains the results of all the SCIO measurements, plus additional feature information for each measurement. The latter contains the results from the DXA measurements.

Feature Name	Description	Values
folio	Woman ID	Numeric
mama	Breast identifier	right, left, left2
ubicacion	Location	3pm, 6pm, 9pm, 12pm
bmi	Body Mass Index	Numeric
fitzpatrick_color	Skin color classification	II, III, IV, V
copa_sosten	Bra size	A, B, C, D
edad	Age	Numeric
spectrum 0-330	SCIO spectrum measurements	Numeric

Table 1: SCIO Table column information

Feature Name	Description	Values
folio	Woman ID	Numeric
mama	Breast identifier	right, left, left2
dxa_density	Breast density measured using DXA	Numeric

Table 2: DXA Table column information

After merging both tables, we have

- 197 folios.
- 3 breast per folio.
- 5 locations per breast.
- 3 spectrum measurements per location

Variability within left breast measurements

We examine now how much does the variability of the SCIO measurements can be attributed to differences in the breast locations, and how much is due the imprecisions of the instrument. The following histogram illustrates the percentage variation between the two measurements of the left breast for each folio, and the percentage variation of a folio with respect to the average.

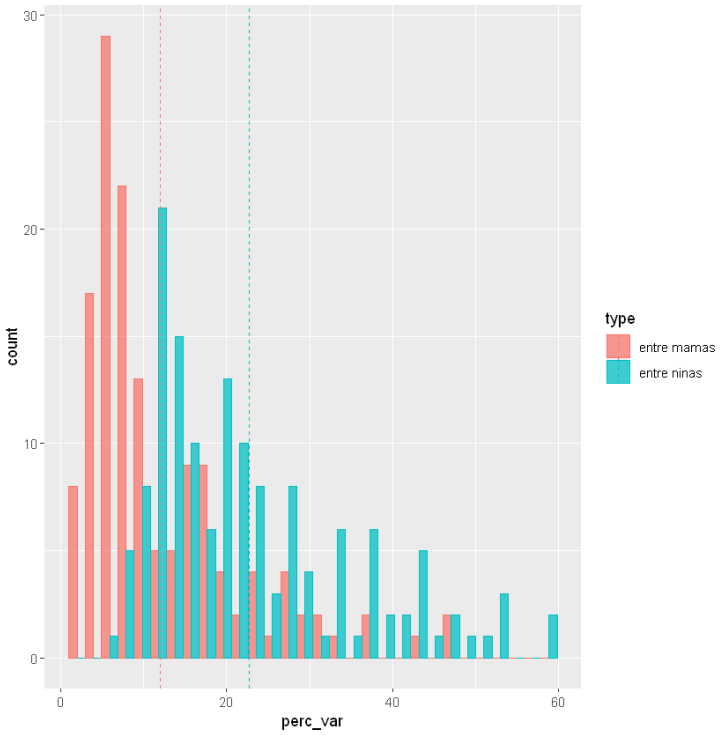

Figure 1: Variability of measurements across women and across breasts for the same woman.

Around 30% of folios have a left breast variation that is indistinguishable from the variation between folios. We decided to remove those entries from our analysis and predictions.

Distribution Analysis

We begin by analyzing the distribution of DXA density values in Figure 2. The blue dashed lines correspond to the 25%, 50% and 75% quantiles.

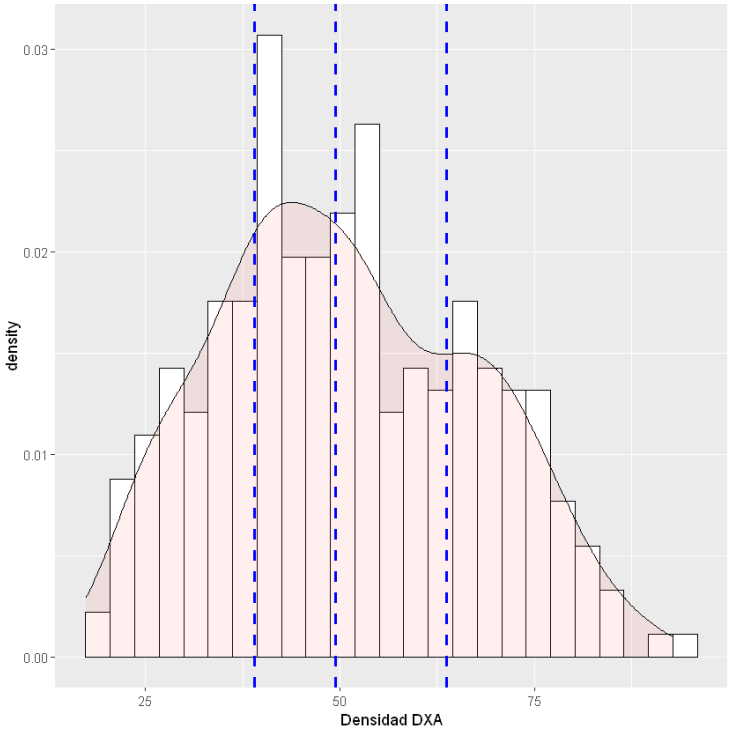

Figure 2: Distribution of DXA density.

Then we analyze the distribution of spectrum values per breast separated by BMI. As a way to illustrate differences in density, we color code the entries with high DXA density in red (75% quantile) and low DXA density in blue. We perform the same analysis for Copa Sosten and Fitzpatrick Color, and we also consider the mean values within each category.

We conclude from this preliminary analysis that the spectrum variables, when controlling for some of these categorical features, can slightly differentiate high from low density individuals.

Another fact to highlight is the autocorrelation of the spectrum series. Figure 9 shows the average correlation an spectrum column (from 0 to 330) has with the $k-$the previous spectrums, for k from 1 to 25. We see that these values are high (from almost 1 with $k = 1$, an above 0.85 for $k = 25$. This suggest that looking at all columns of the spectrum in the analysis is redundant.

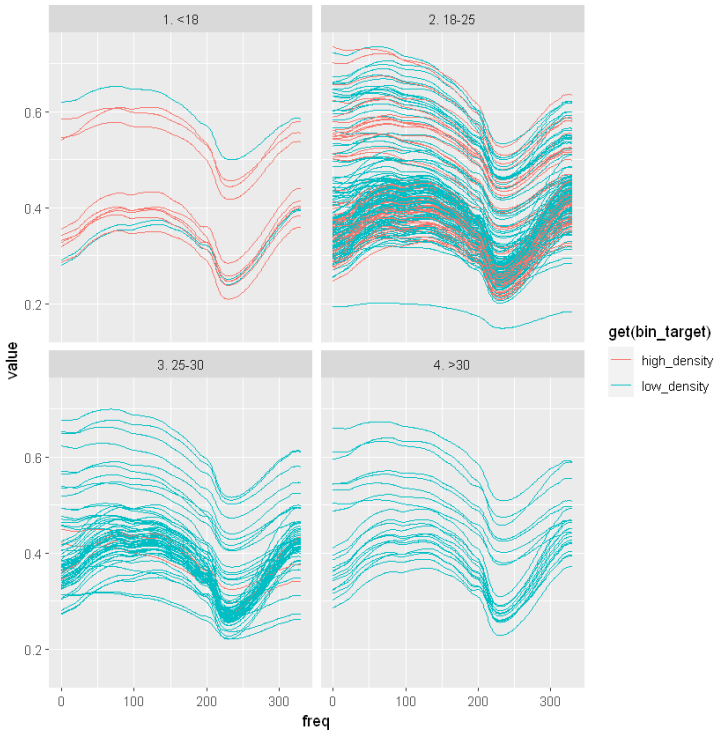

Figure 3: Distribution of spectrum values (for index 0 to 330), by different BMI categories.

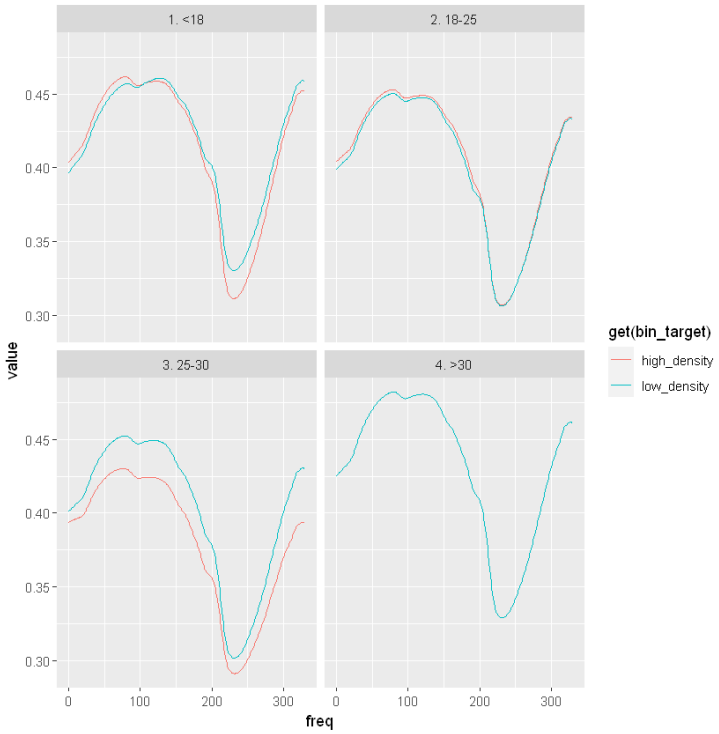

Figure 4: Mean of spectrum values (for index 0 to 330), by different BMI categories.

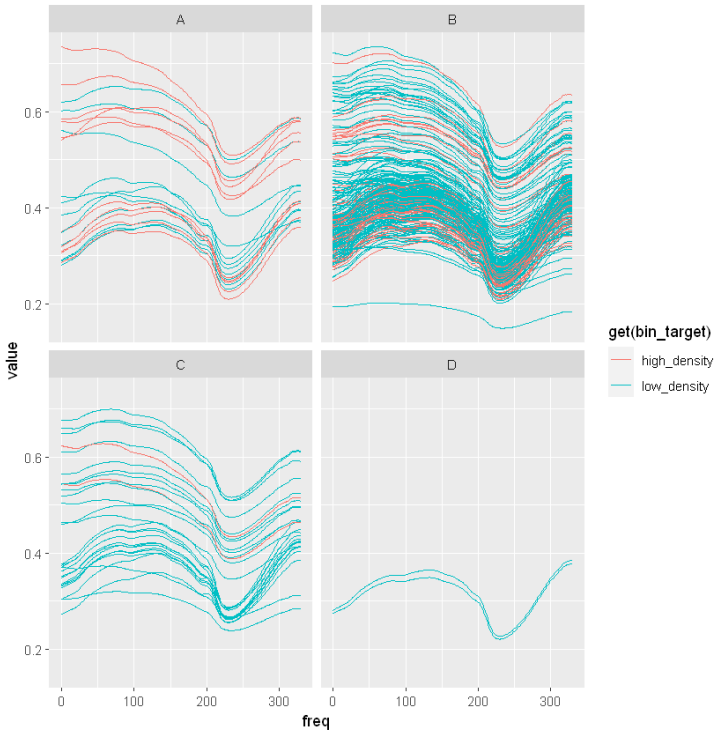

Figure 5: Distribution of spectrum values (for index 0 to 330), by different BMI categories.

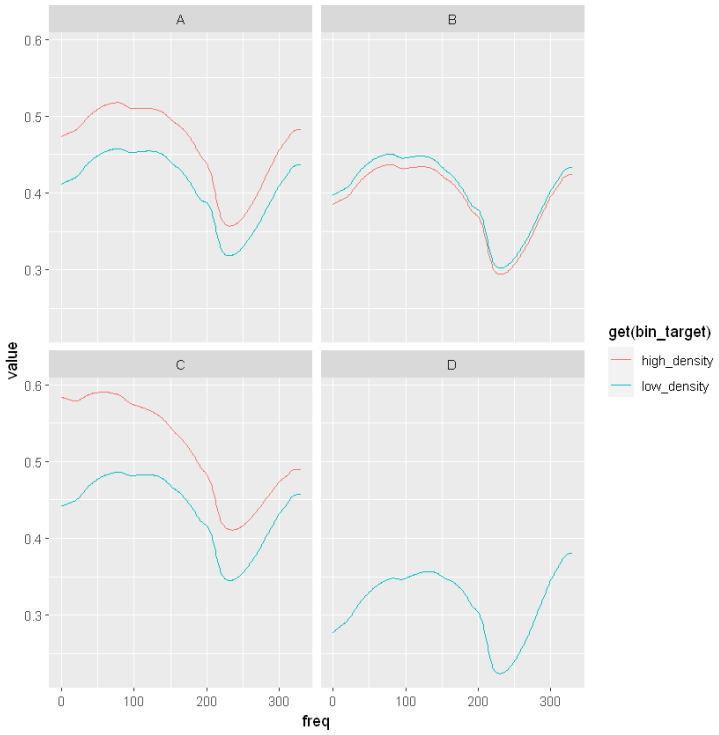

Figure 6: Mean of spectrum values (for index 0 to 330), by different BMI categories.

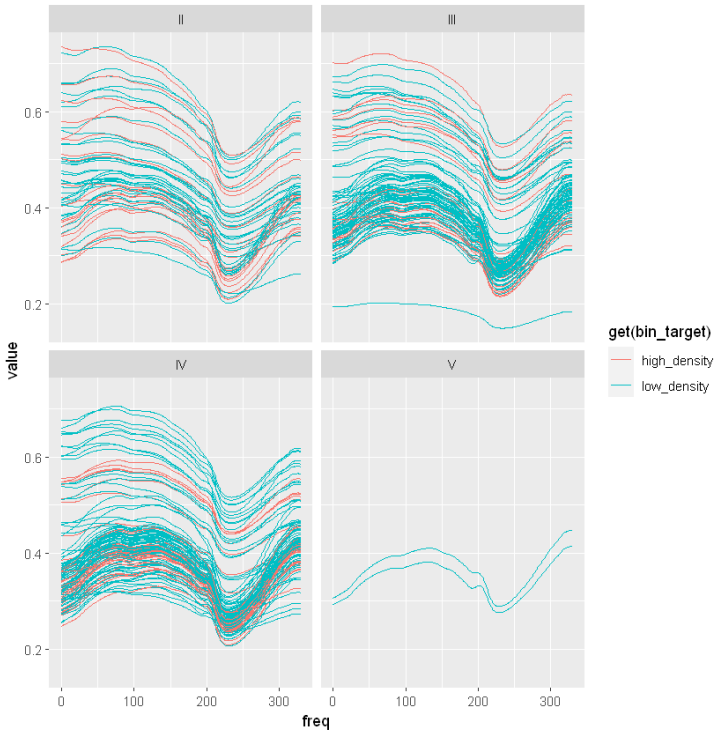

Figure 7: Distribution of spectrum values (for index 0 to 330), by different BMI categories.

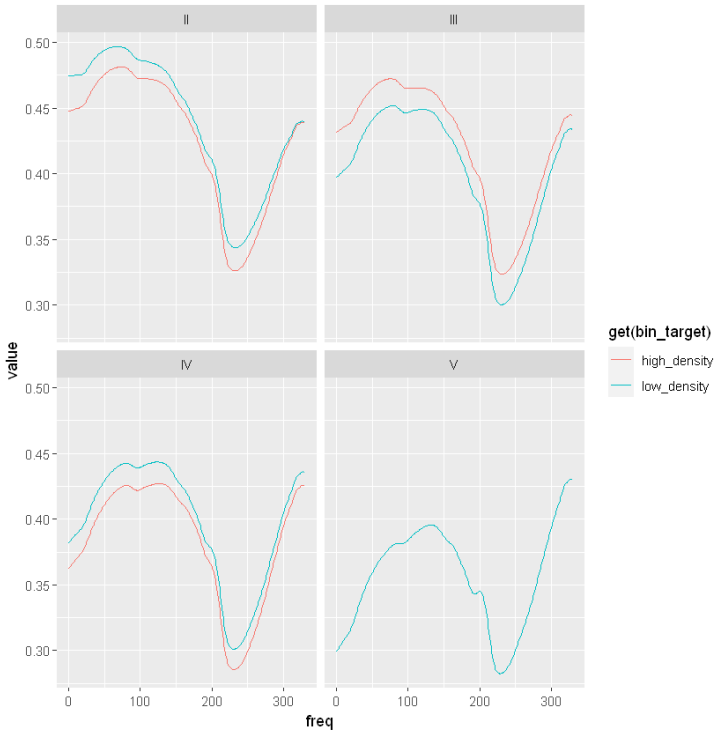

Figure 8: Mean of spectrum values (for index 0 to 330), by different BMI categories.

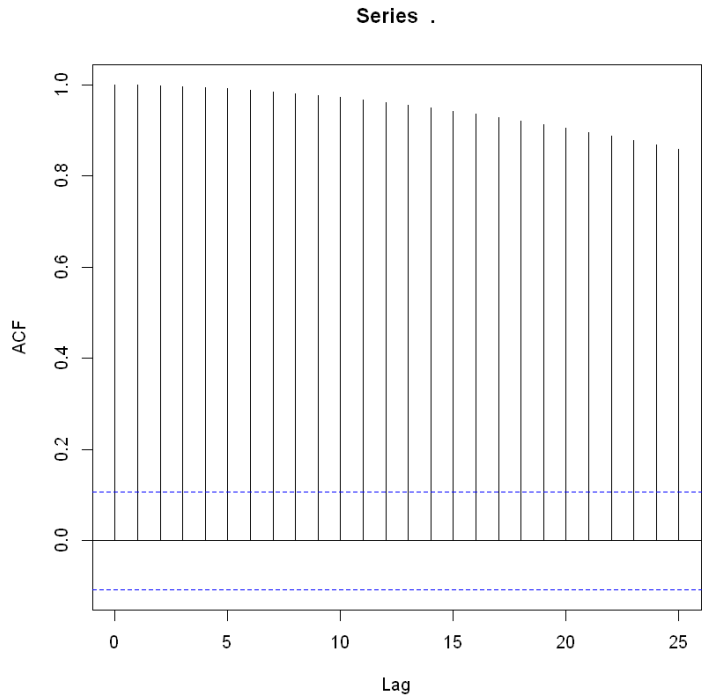

Figure 9: Autocorrelation of spectrum values.

Prediction

Now we show the results of our classification prediction experiments. We consider as features the numeric BMI, age, Fitzpatrick color, and bra size. To aggregate spectrum, we took moving averages of size 50 and overlap of 25. This decision is taken based on the autocorrelation analysis, but the values of the window and overlap is somewhat arbitrary. As a target we considered the binary variable of high or low density, set at whether the DXA density is at value above or below 60 (which is roughly the 75% quantile in the sample).

We built predictors using four different methods: logistic classification, support vector machine classifier, random forests, and optimal trees with hyperplanes. We compared the accuracy of our model – and AUC, when available – with that of the naive benchmark (predict largest category) and the model that does not consider spectrum as a feature.

Figure 10: Accuracy of each model on test data (out of sample accuracy).

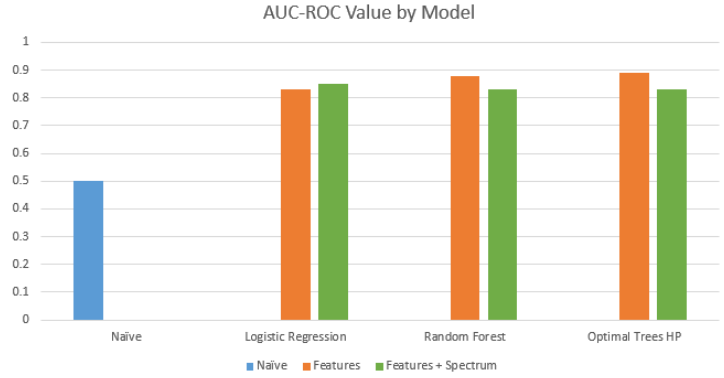

Figure 11: Accuracy of each model on test data (out of sample accuracy).

Both models outperform significantly the naive model. There exists some win in accuracy adding the spectrum values to the model, although it is not that significant. However, the difference grows slightly more than what is shown in the graph the smaller the moving average window becomes.

Next, we show the confusion matrix for each of this models. We compute the number of correct and incorrect label predicted for each class

Finally, we give an analysis based on the feature importance to show that although the spectrum is not the most relevant set of features, they are not negligible. TODO

Predicted/Actual	Low Density	High Density
Low Density	66	24
High Density	4	28

Table 3: Logistic classification confusion matrix.

Predicted/Actual	Low Density	High Density
Low Density	72	18
High Density	3	29

Table 4: SVM confusion matrix.

Predicted/Actual	Low Density	High Density
Low Density	65	25
High Density	3	29

Table 5: Random Forest confusion matrix.

Predicted/Actual	Low Density	High Density
Low Density	82	8
High Density	4	24

Table 6: Optimal Tress with HP confusion matrix.