Exploring SHAP Values in Imbalanced: Insights on Bias and Concept Drift

Antonio Guillén Teruel

UNIVERSIDAD DE MURCIA

Topics

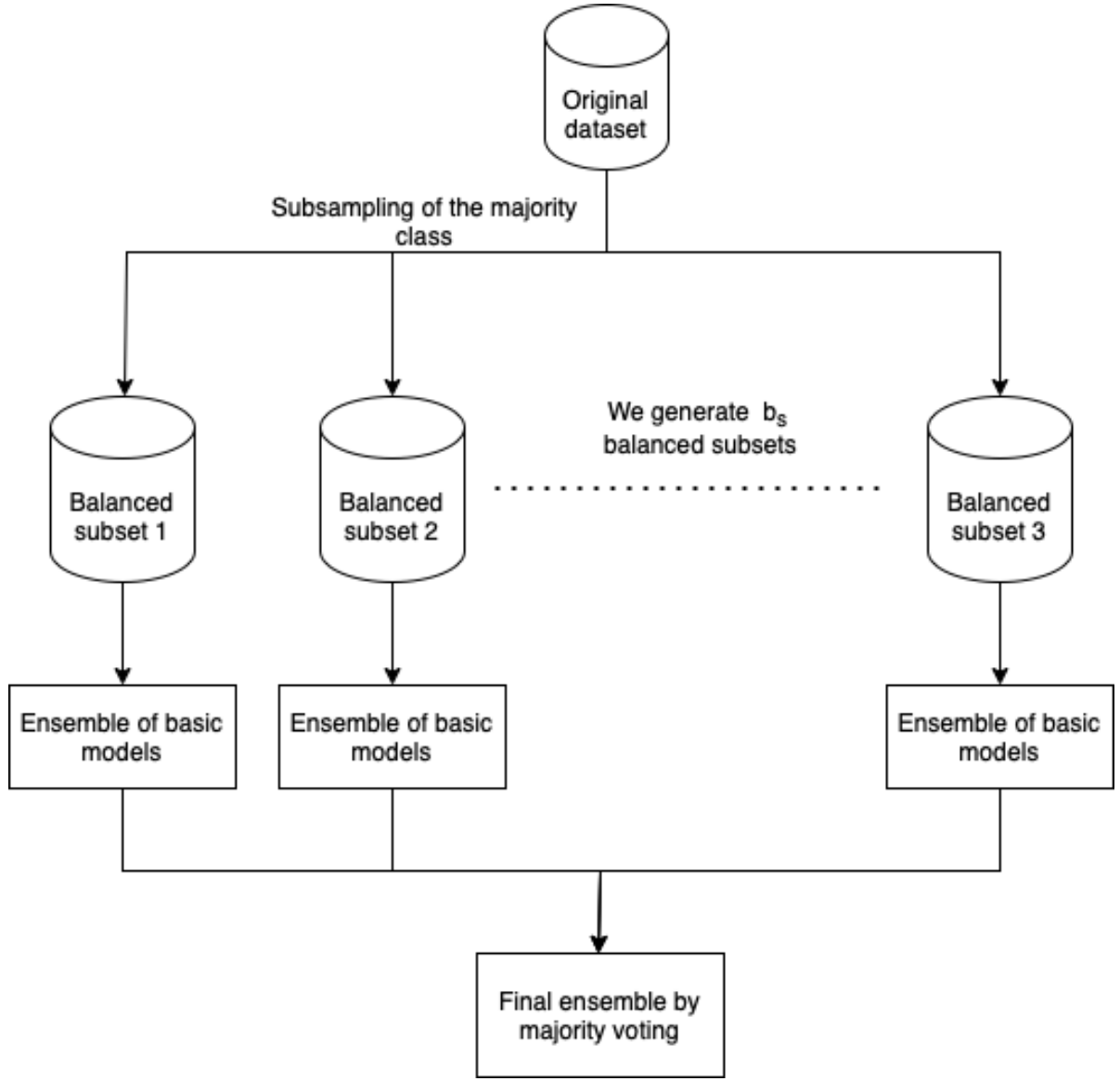
- Previous related work
  - IPIP
- 3 months stay in Krakow
  - IPIP+CD: Learning in Imbalanced Problems
  - Explainability perspective of CD with IPIP
  - Bias in SHAP values by the Imbalanced
- Next steps - To do

Context

- We're working with binary classification tasks.
- Our problems are imbalanced: Minority and majority class.
- Focus on :
    - Concept Drift and identifying it using SHAP values.
    - Effect of the imbalanced in the explainability of the models/data: SHAP values.

IPIP

- "b_s" balanced subsets (55%-45%) are generated by subsampling the majority class so that all elements of the minority class are represented in at least one of the subsets.
- An ensemble is trained for each subset. Models are added by majority voting if the previous ensemble is improved.
- Finally, the final prediction is what the majority of ensembles decide.



Concept drift: Passive learning

- We have a time stamped dataset. If X are the predictor variables of our data and Y the target variable, we call a concept:

  Concept=P(X,Y)

- So there is a concept drift when:

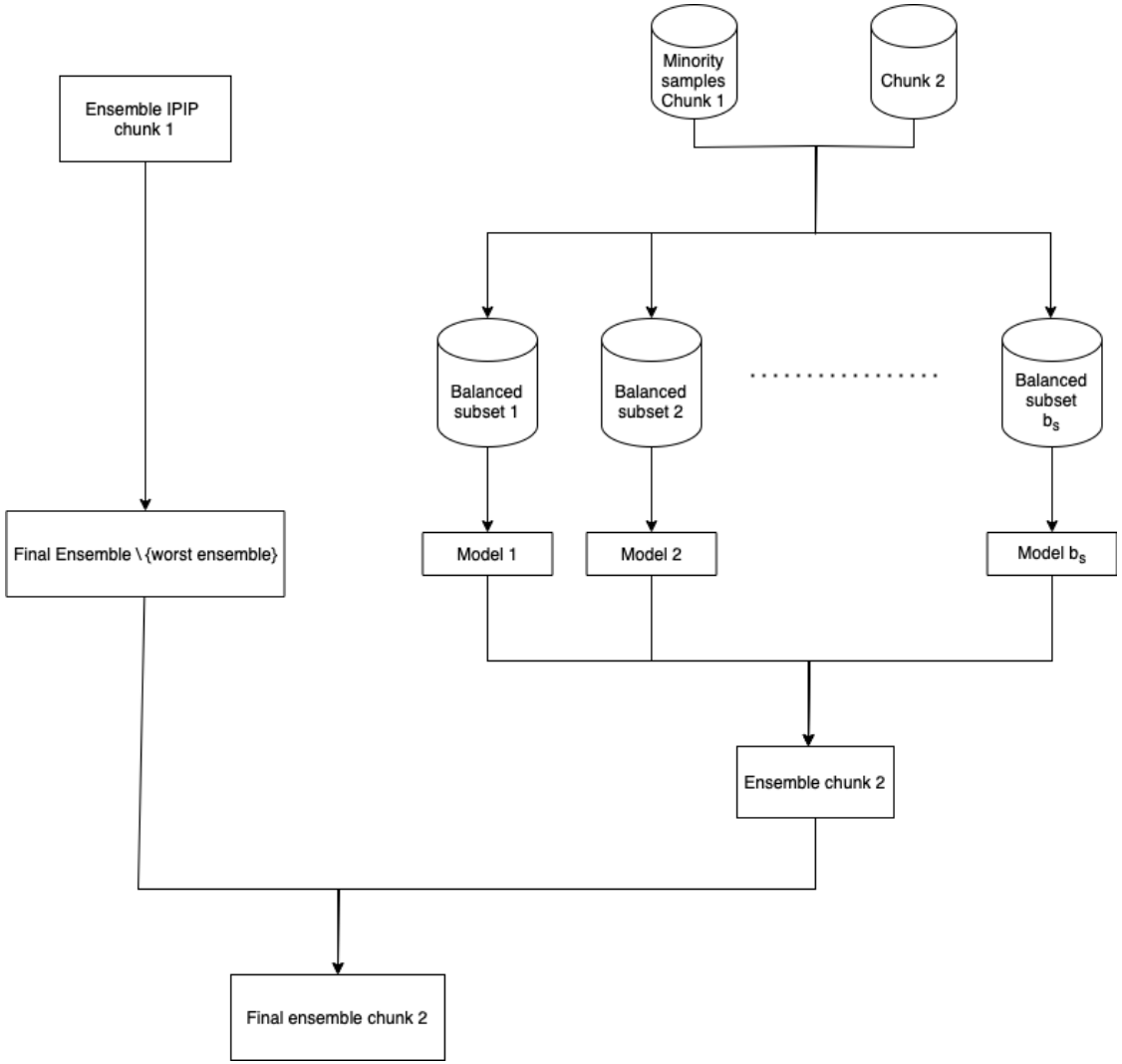
$$P_t(X,Y) \neq P_u(X,Y)$$

  So that *t* and *u* are different time stamps.

- Rather than attempting to identify a concept drift, we will assume that data evolves over time and adapt the model accordingly. Then, we are dealing with passive learning. It is even more challenging when dealing with imbalanced data.

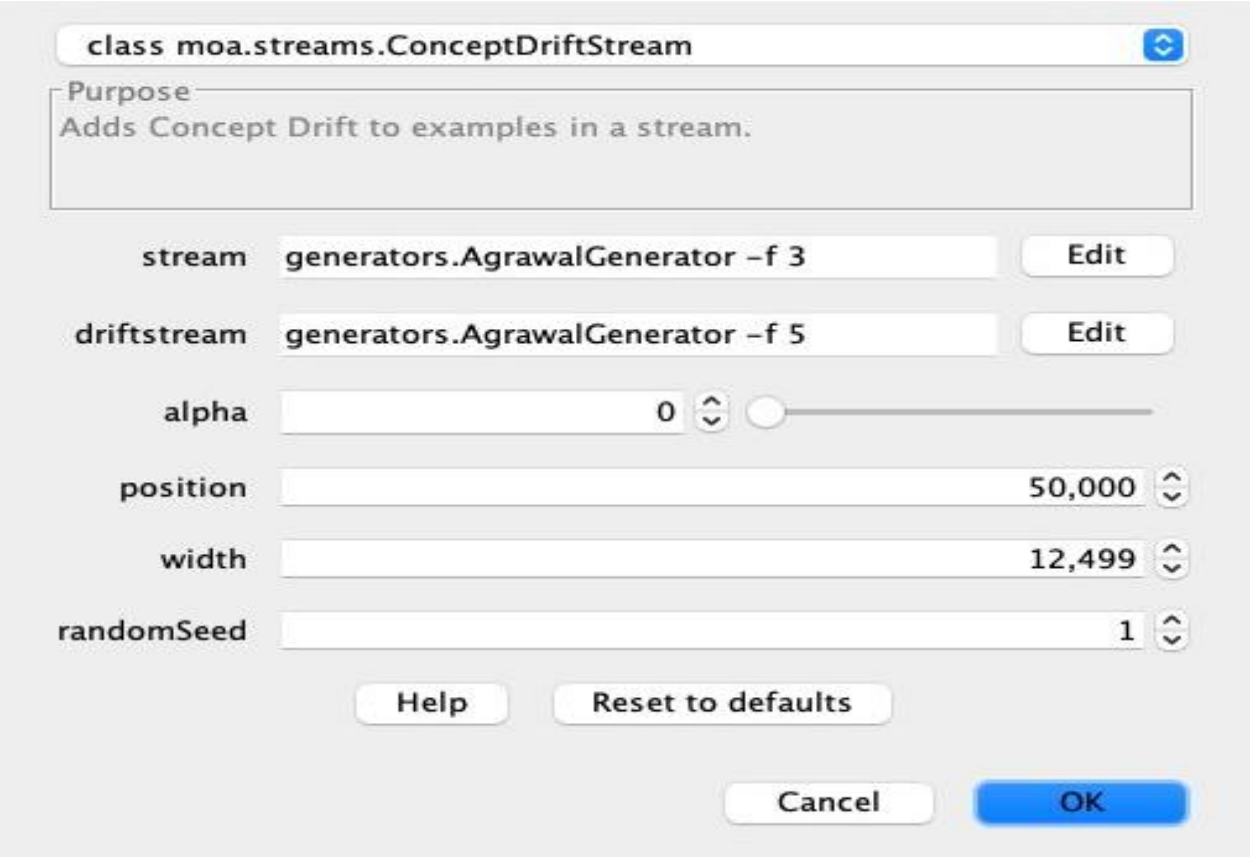
- The IPIP method is ensemble-based, thereby facilitating adaptation to the needs of a passive learner.

IPIP + Concept Drift



- Let's assume that data is divided by chunks. We want to update IPIP with each new chunk of data.
- Firstly, we train a basic approach of IPIP with the first chunk of data.
- Now, for next chunks the approach is to update the previous IPIP final ensemble.

Datasets for the experiments



1. We use MOA (Massive Online Analysis).
2. Datasets are simulated with the drift of your choice (Abrupt or Gradual).
3. Preprocessing is performed to separate the total dataset into 34 chunks of 1000 instances with an unbalance of 80%-20%.
4. Drifts after chunks 13 y 25.
5. IPIP is trained with those chunks.
6. Is it possible to visualise a change in SHAP values over the course of the chunks?

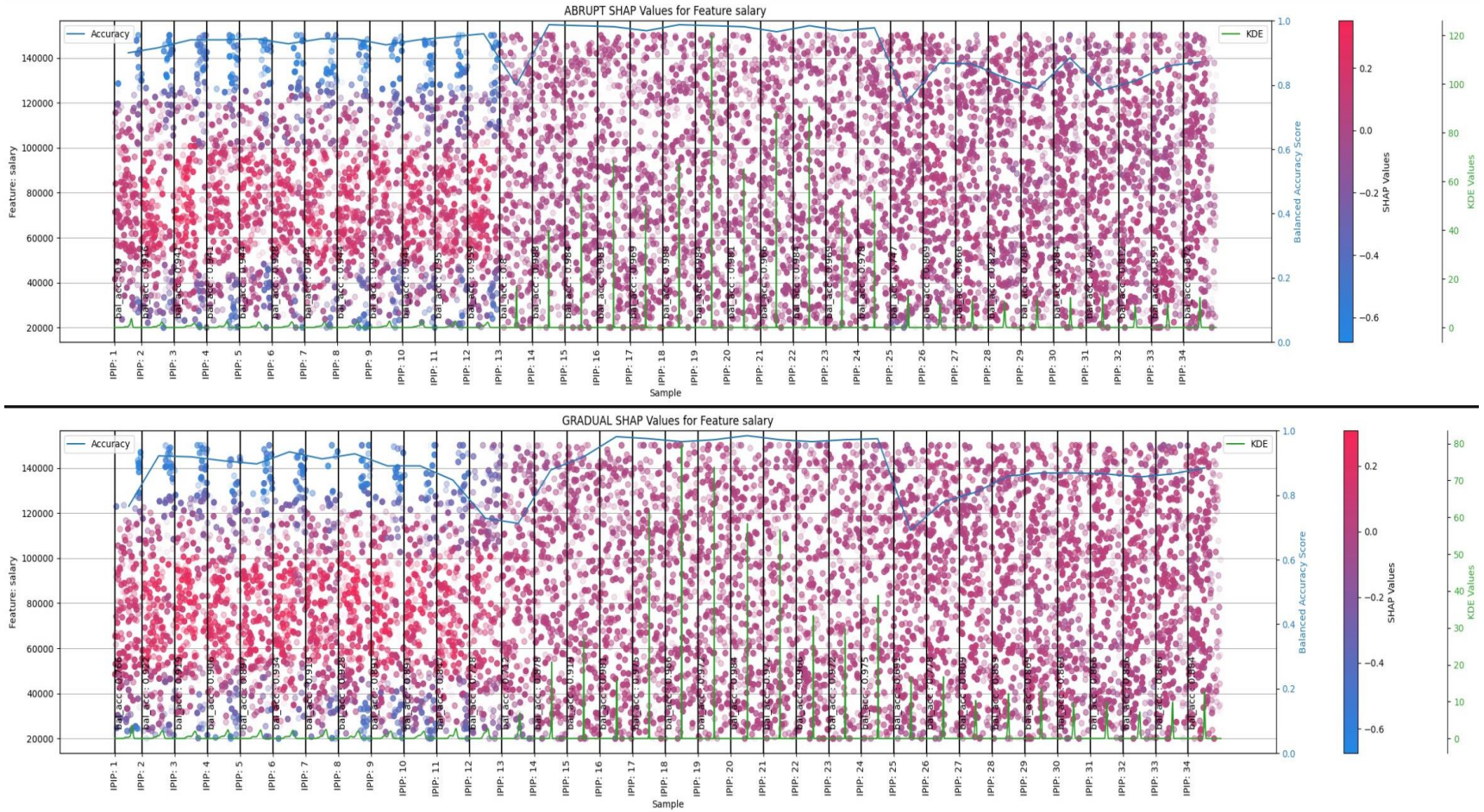
Triangle of explainability: Data, model and explanations

- We want a method that will allow to spot selected types of concept drifts in streaming imbalance data through:
    - **Data dimension**: What's the distance of the data distribution between chunks?
    - **Model dimension**: We can use the model (IPIP) confidence as an indicator of possible drift.
    - **Explanation dimension**: We have a method for calculating SHAP values from IPIP (Natalia's work) over time. We can then analyse SHAP values.

The idea is to append all this information to identify each type of drift.

AGRAWAL synthetic dataset results

- **Temporal data**: SHAP values are obtained for each instance of the test sets for each chunk over time.
- **Colour trend**: An attempt is made to see if there is a trend or a change in trend over time for some variables.
- **Kernel Density Estimation**: A plot of the distribution of the SHAP values is used to observe changes.
- **Model performance**: The Balanced Accuracy of each chunk is displayed to see the effect of drift.



Next steps

Triangle of explainability work:

- Different datasets and different drifts to try with IPIP + SHAP values.
- Different models to compare with IPIP results.
- Perform numerical analysis to identify the drift in a non-visual way.
- Can we identify some drift that can't be determined by looking at the model performance?

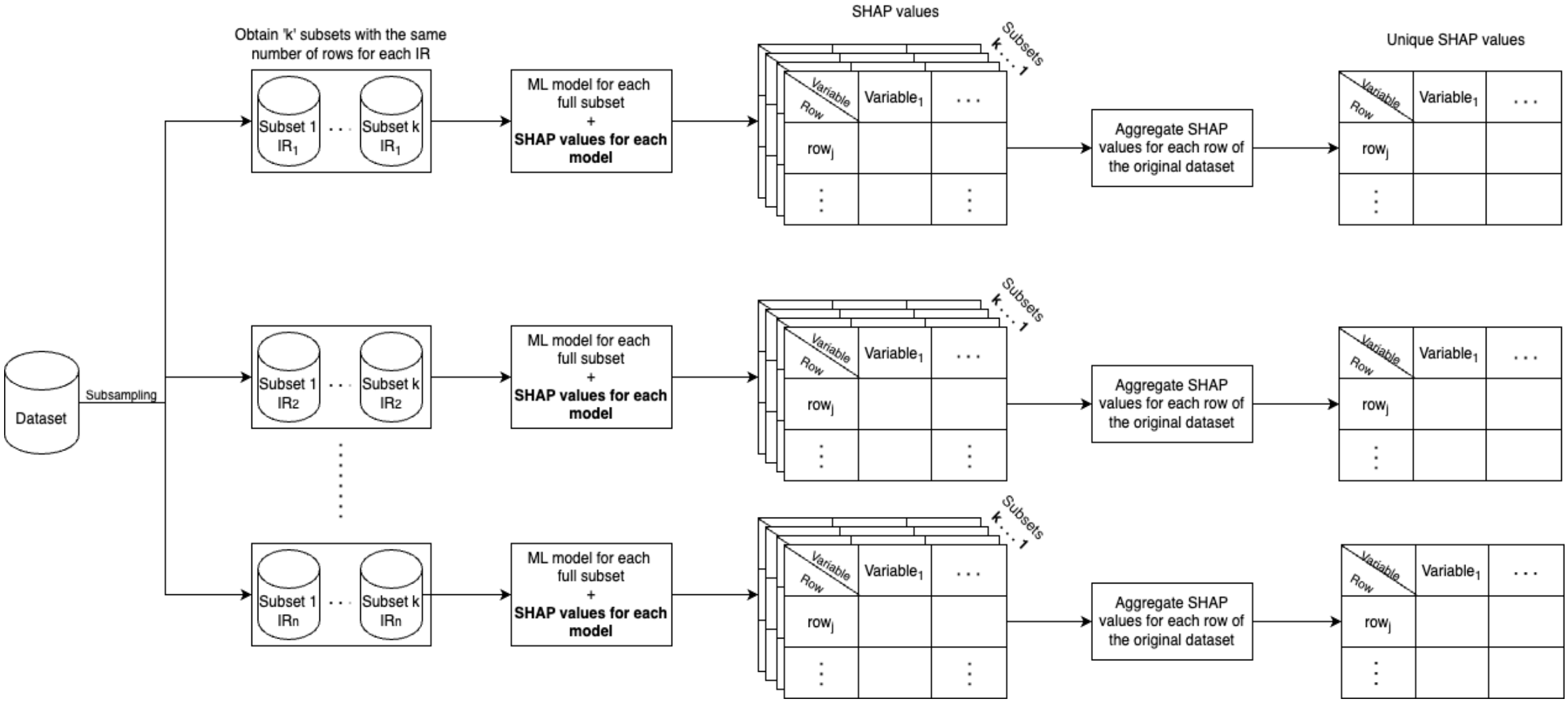
Bias in the SHAP values by an imbalanced problem

How much are SHAP values affected by the imbalance of a binary classification problem?

Experiments setup:

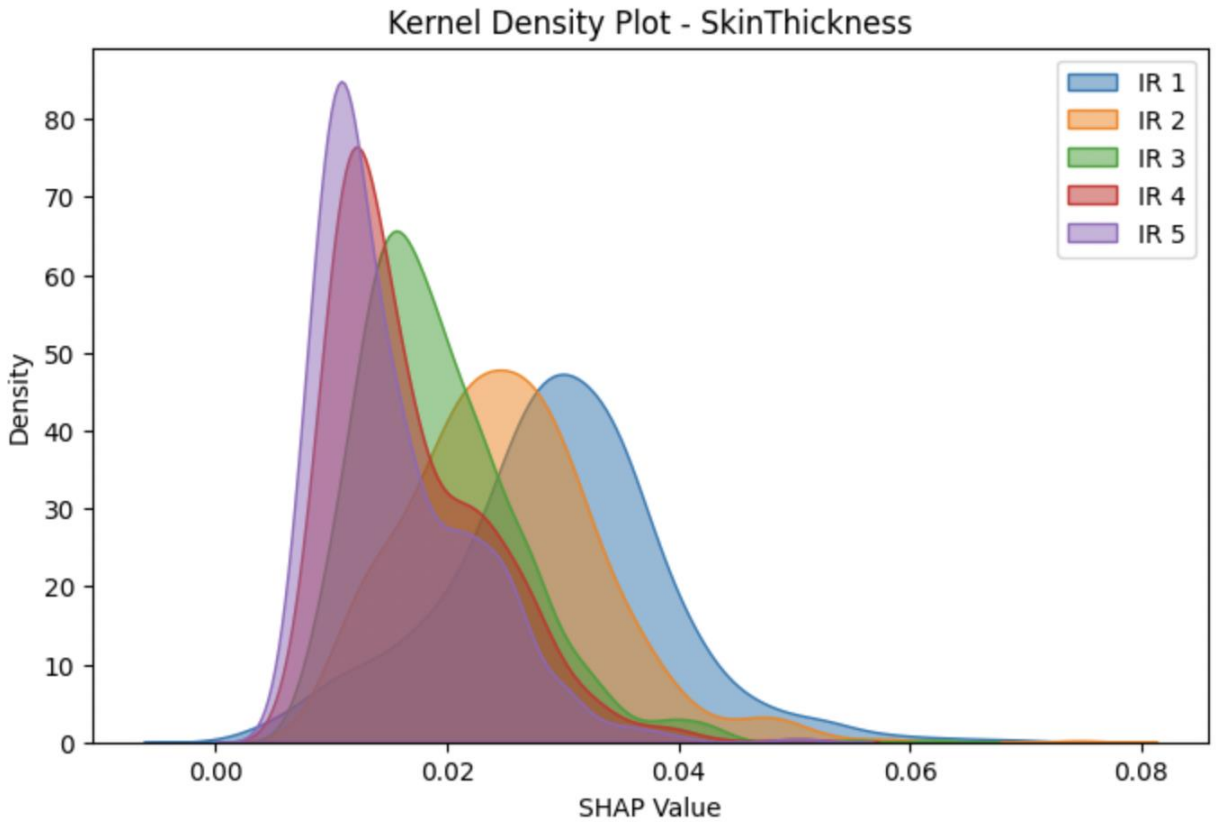
- Model: Random Forest
- 5 imbalanced datasets for a variety of results: COVID, Pima, Satimage, Forest and Mammography.
- Each dataset will be sampled into subsets with a specific Imbalanced Ratio (IR).
- 5 different Imbalanced Ratios, with the associated minority class proportion: 2%, 5%, 10%, 15%, 20%
- Then we will analyse the effect of the different IRs in the SHAP values.

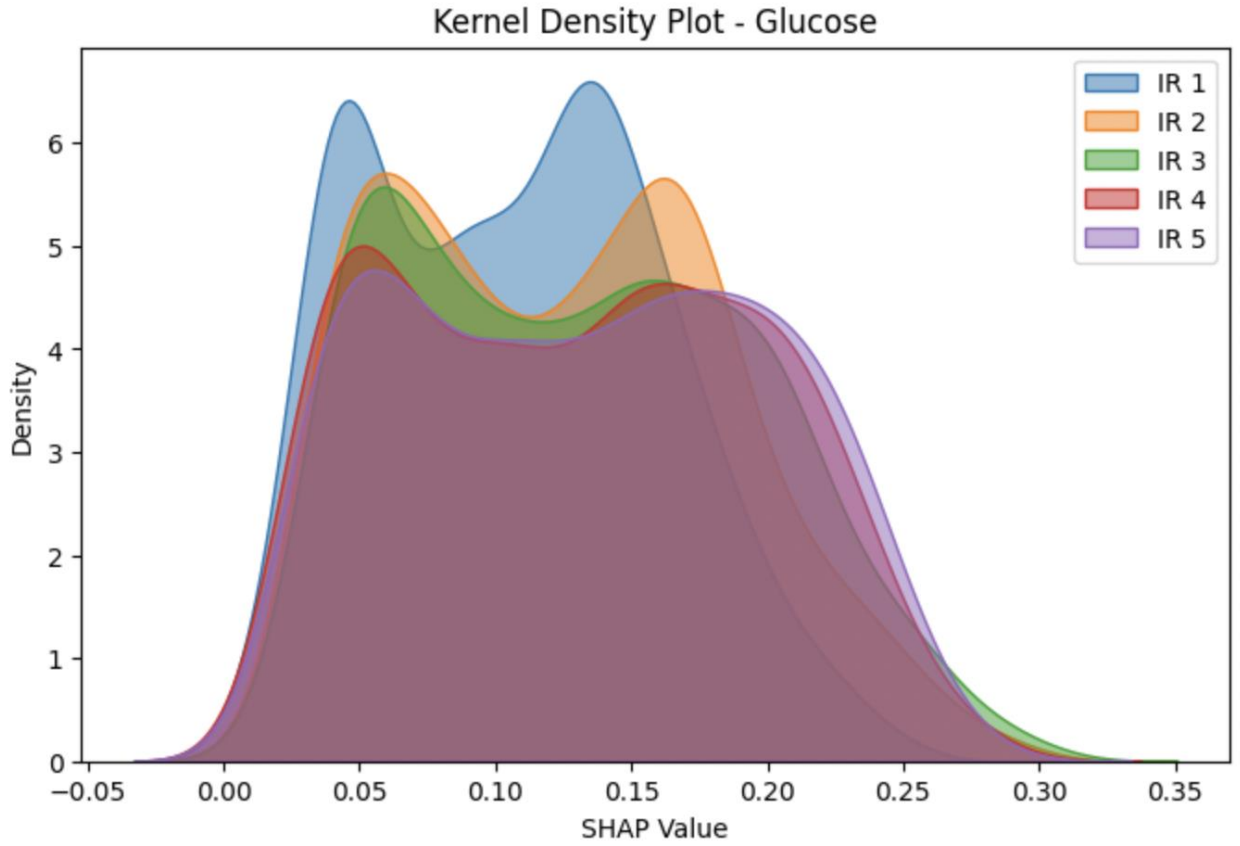
How can we obtain the subsets for each IR and the SHAP values?



Analysis performed with the SHAP values: Pima

1. Kernel Density Plots of the SHAP values distribution for each variable.





Analysis performed with the SHAP values: Pima

2. Measures as the mean, median, SD or IQR of SHAP values per variable and IR.

Variable	IR	Mean	Median	SD	IQR	Mean_Rank	SD_Rank	IQR_Rank	Median_Rank
Pregnancies	1	0.039318	0.038882	0.015402	0.021417	1	2	4	1
Pregnancies	2	0.035818	0.033173	0.015288	0.022621	2	3	2	2
Pregnancies	3	0.032346	0.03072	0.014432	0.019818	4	5	5	5
Pregnancies	4	0.032249	0.031596	0.015189	0.022374	5	4	3	4
Pregnancies	5	0.032562	0.033115	0.016673	0.024359	3	1	1	3

Analysis performed with the SHAP values: Pima

3. Correlation between measures as the mean, median, SD or IQR of SHAP values and IRs

Variable	Metric	Correlation	P-Value
Glucose	IQR	0.913865	0.029951
Glucose	Median	0.87905	0.049568
BloodPressure	Mean	-0.93558	0.019437
BloodPressure	SD	-0.949738	0.013424
BloodPressure	IQR	-0.944536	0.015549
BloodPressure	Median	-0.913114	0.03034
SkinThickness	Mean	-0.955967	0.011018
SkinThickness	SD	-0.89328	0.041174
SkinThickness	Median	-0.959249	0.009815
Insulin	Mean	-0.905918	0.034148

Next steps

Bias in SHAP values in imbalanced problems:

- Perform statistical tests to see significant differences between distributions of SHAP values (KS-test, ANOVA, …)
- Perform more complex analysis as getting Rand Index for the set of most important features for the subsets of each IR: Association rules.
- Compare the results of the Rand Index between IRs.

Thank you! :-)