

# Concept drift in imbalanced problems

Antonio Guillén-Teruel

José Palma

Juan A. Botía



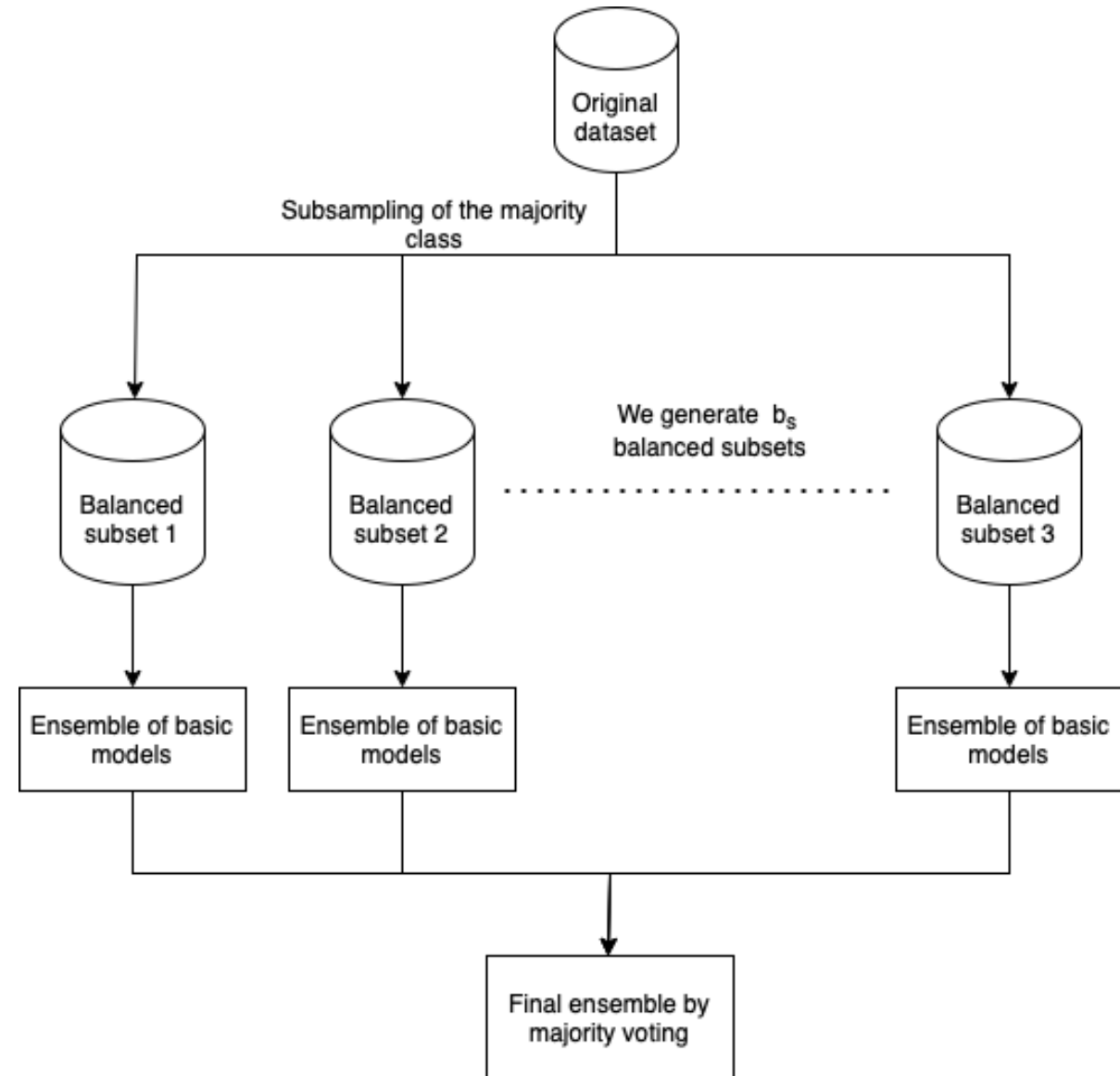
UNIVERSIDAD  
DE MURCIA

# Topics

- Previous work
  - IPIP (Identical Partitions for Imbalance Problems)
  - UIC
  - Application of IPIP in a real case
- Work in progress – Concept Drift (CD)
  - Passive CD
  - IPIP+CD
  - R Package with 8 passive learners
  - Some results
- Future work

# IPIP (Identical Partitions for Imbalance Problems)

- " $b_s$ " 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.
- Paper under review



# UIC

We have an unbalanced dataset 'd'.

We obtain several subdatasets by sampling the original dataset varying the imbalanced ratio, training and evaluating models in each of them.

The idea is that we can create a new measure, as an integration of all the biased ones, by means of an aggregation in which all the basic measures are inversely weighted by their respective correlation with the proportion of the minority class (IR) of each dataset, with the intention that UIC is less biased with the IR.

R package available (paper under review)

<https://github.com/antoniogt/FILM>

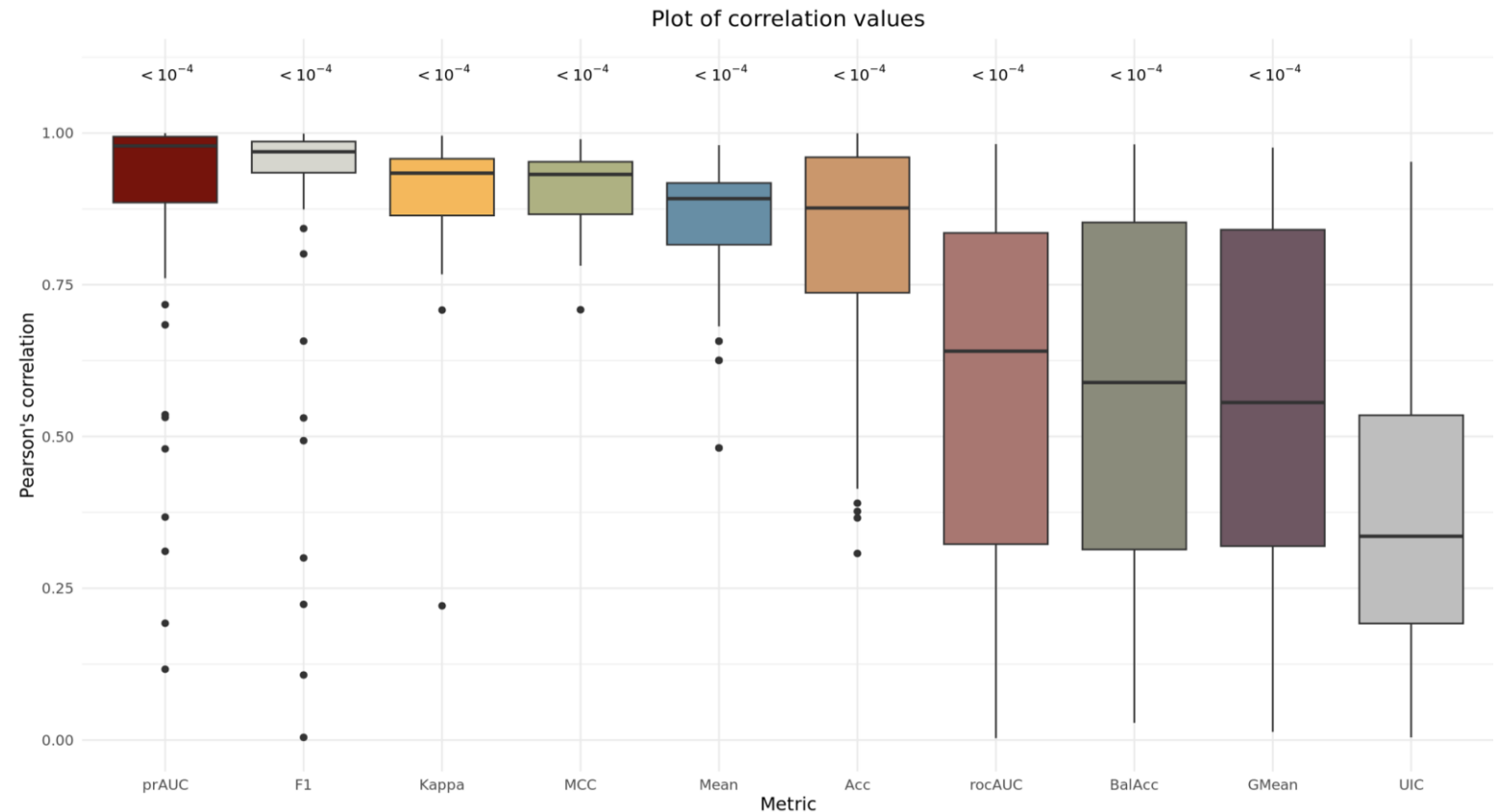


Figure: Pearson's Correlation between some basic metrics, UIC and the Imbalance Rate


## Our work to address imbalance in a real COVID-19 patient cohort classification (published)

- We applied IPIP: We achieved a balanced accuracy near 93% and a ROC-AUC of 0.94 by testing IPIP on a validation set to predict the final condition of a COVID-19 patient using comorbidities and demographic data

[nature](#) > [scientific reports](#) > [articles](#) > [article](#)

Article | [Open access](#) | [Published: 28 October 2022](#)

### **A predictive model for hospitalization and survival to COVID-19 in a retrospective population-based study**

[Alejandro Cisterna-García](#), [Antonio Guillén-Teruel](#), [Marcos Caracena](#), [Enrique Pérez](#), [Fernando Jiménez](#), [Francisco J. Francisco-Verdú](#), [Gabriel Reina](#), [Enrique González-Billalabeitia](#), [José Palma](#), [Álvaro Sánchez-Ferrer](#) & [Juan A. Botía](#) 

[Scientific Reports](#) **12**, Article number: 18126 (2022) | [Cite this article](#)

**2824** Accesses | **8** Citations | **23** Altmetric | [Metrics](#)

#### **Abstract**

The development of tools that provide early triage of COVID-19 patients with minimal use of diagnostic tests, based on easily accessible data, can be of vital importance in reducing COVID-19 mortality rates during high-incidence scenarios. This work proposes a machine learning model to predict mortality and risk of hospitalization using both 2 simple demographic features and 19 comorbidities obtained from 86,867 electronic medical records of COVID-19 patients, and a new method (LR-IPIP) designed to deal with data imbalance problems. The model was able to predict with high accuracy (90–93%, ROC-AUC=0.94) the patient's final status (deceased or discharged), while its accuracy was medium (71–73%, ROC-AUC=0.75) with respect to the risk of hospitalization. The most relevant characteristics for these models were age, sex, number of comorbidities, osteoarthritis, obesity, depression, and renal failure. Finally, to facilitate its use by clinicians, a user-friendly website has been developed (<https://alejandrocisterna.shinyapps.io/PROVIA>).

# 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= $P(X,Y)$

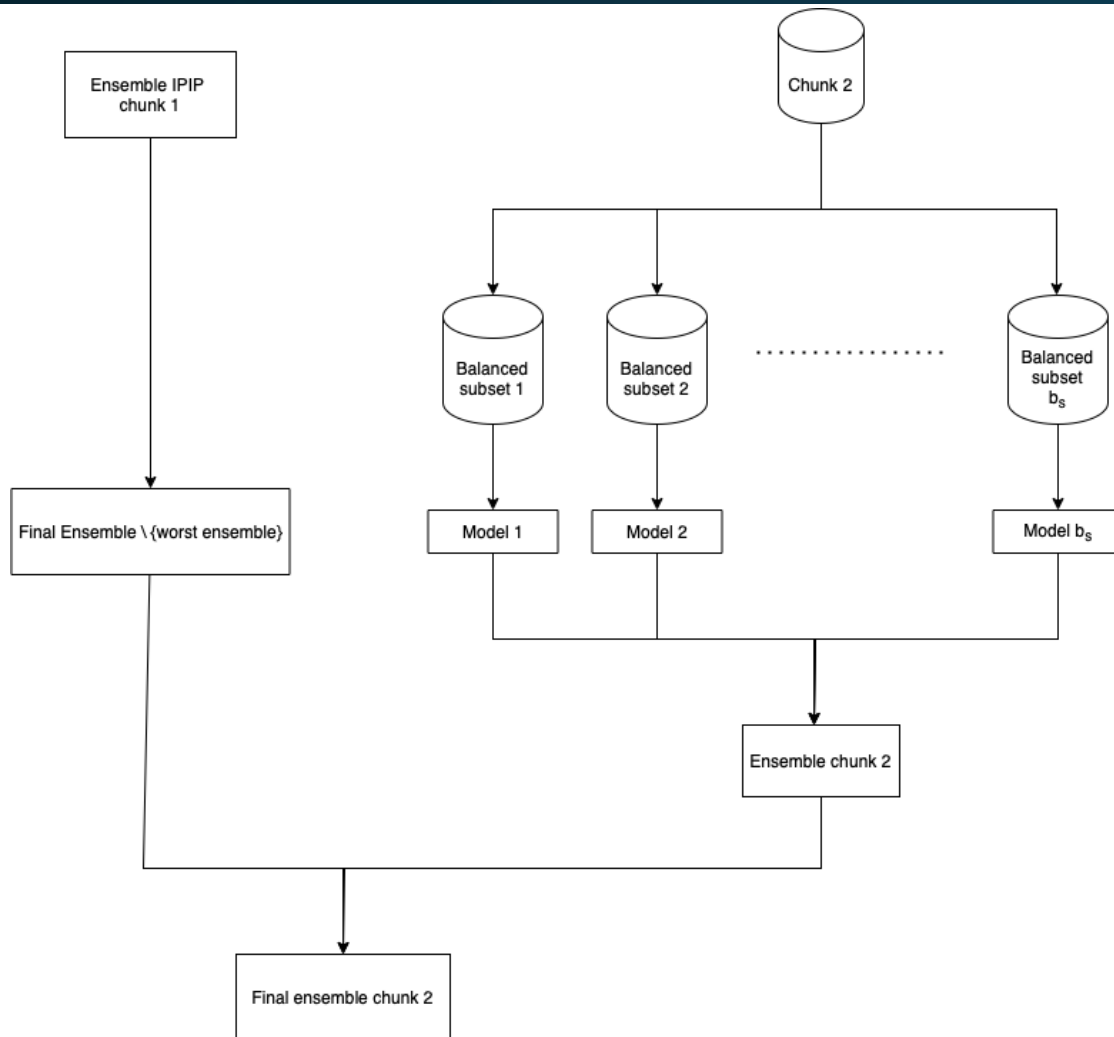
- So there is a concept drift when:

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

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

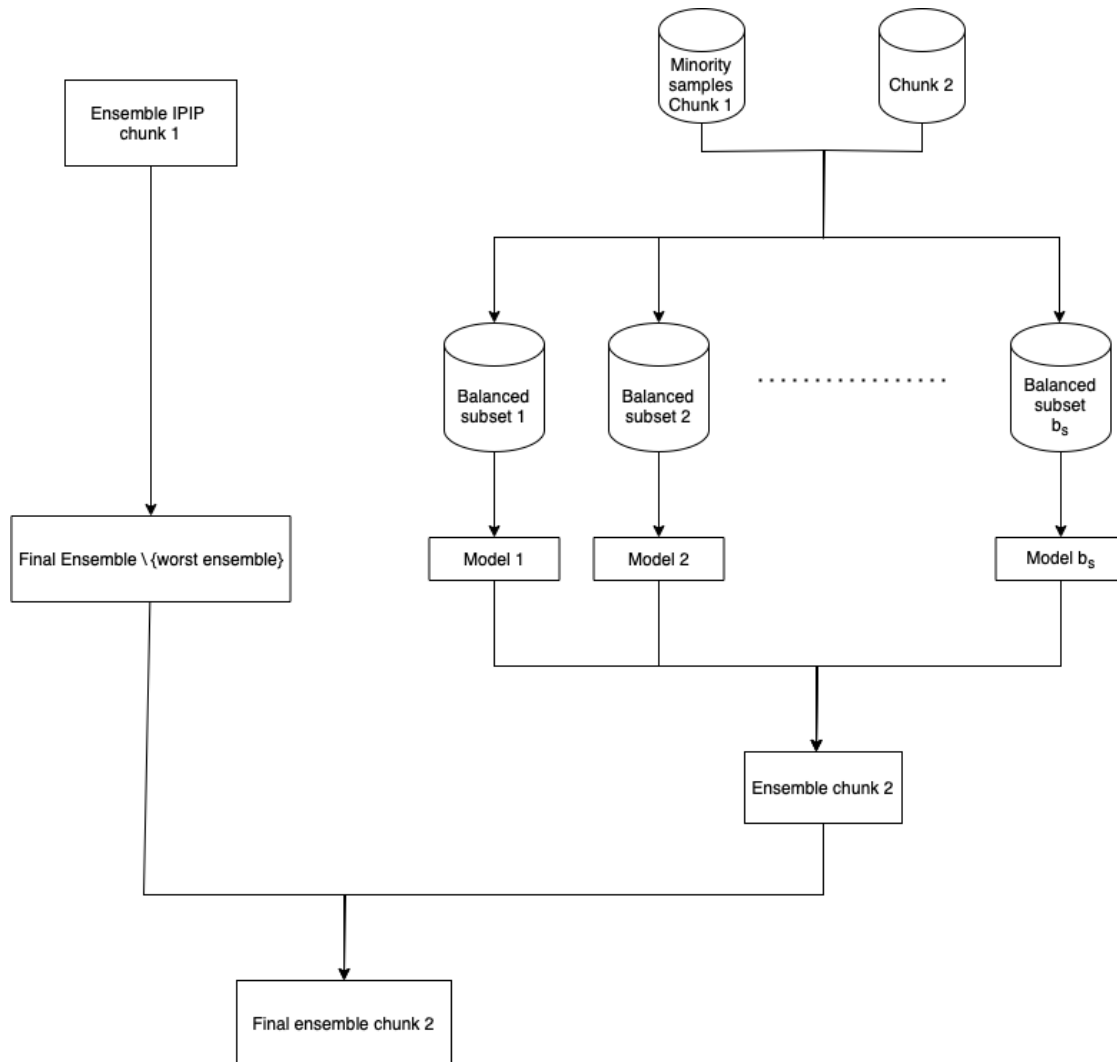
- Rather than attempting to identify a concept drift, we will assume that data evolves over time and adapt the model accordingly. Then, we are dealing with passive learning. It is even more challenging when dealing with imbalanced data.
- IPIP method is ensemble-based, thereby facilitating adaptation to the needs of a passive learner.

# IPIP + Concept Drift Non stationary environment



- Let's presume 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.
- Then, for next chunks the approach is to update the previous IPIP final ensemble with a new ensemble trained with the new chunk of data following the idea of IPIP to train models with balanced subsets and ensuring that every minority class instance is represented in, at least, one subset.

# IPIP + Concept Drift Stationary Environment



- Now, for next chunks the approach is to update the previous IPIP final ensemble with a new ensemble trained with the new chunk of data AND THE PREVIOUS MINORITY CLASS INSTANCES, again following the idea of IPIP to train models with balanced subsets and ensuring that every minority class instance is represented in, at least, one subset.

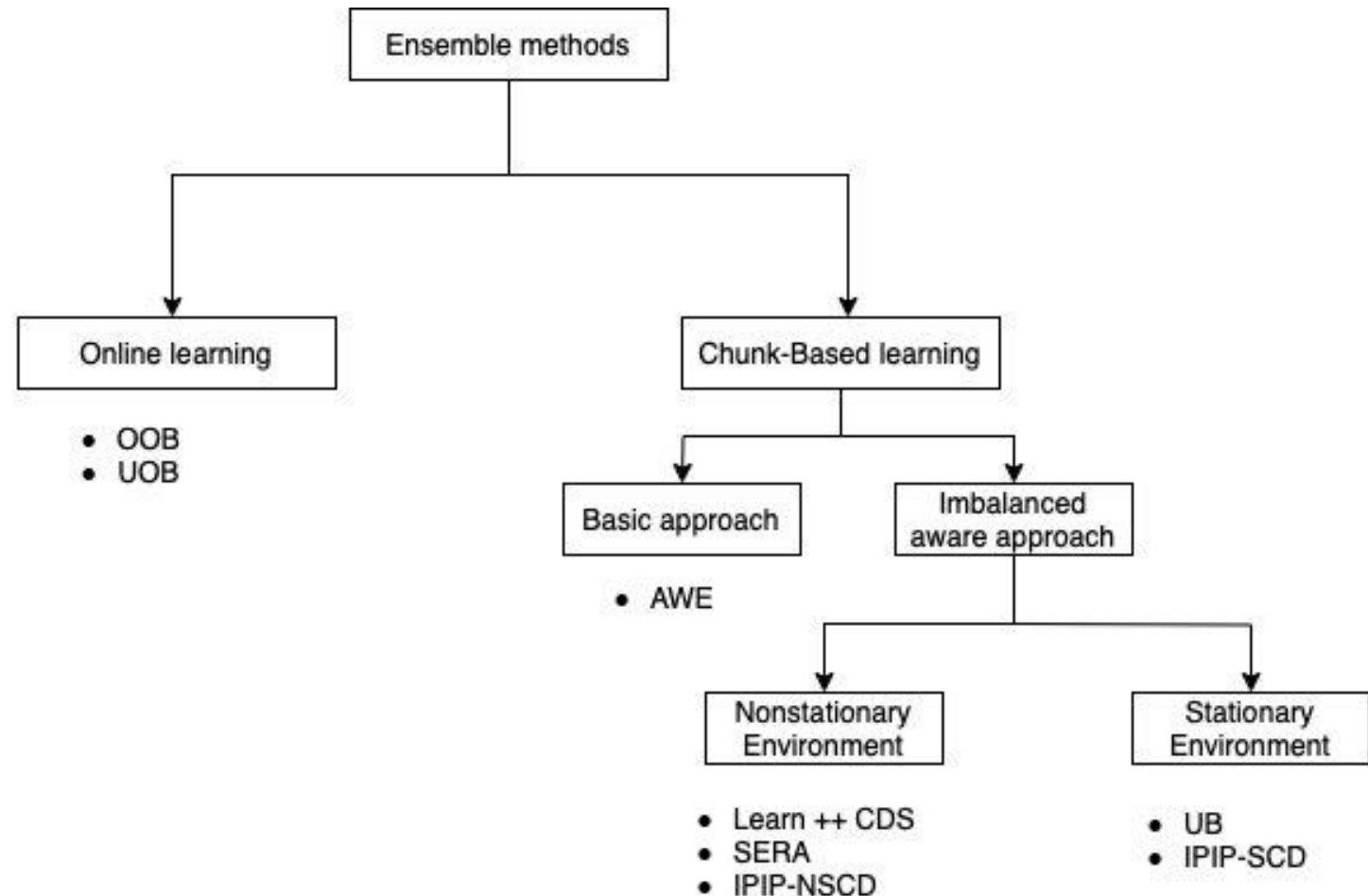


# Our experiments to compare IPIP+CD with other passive learners

- We have some commonly used datasets to study the CD in classification tasks as 'Electricity', 'Weather', and some simulated datasets. We also have a real COVID-19 dataset (SMS) with 48 variables, most of them categorical. COVID-19 data is structured in monthly chunks (26 months) with comorbidities, symptoms, vaccination and demographic data.
- Firstly, we are going to test several ensemble-based passive methods fixing the number of instances per chunk ( $n=1000$ ). And we evaluate each model in the chunk 't' with the chunk 't+1'. We have different experiments artificially varying the IR of the chunks to study the influence of different IRs in a concept drift problem.
- Metrics: Acc, Bal Acc, Kappa, F1, AUC-ROC, AUC-PR, Geom and UIC.

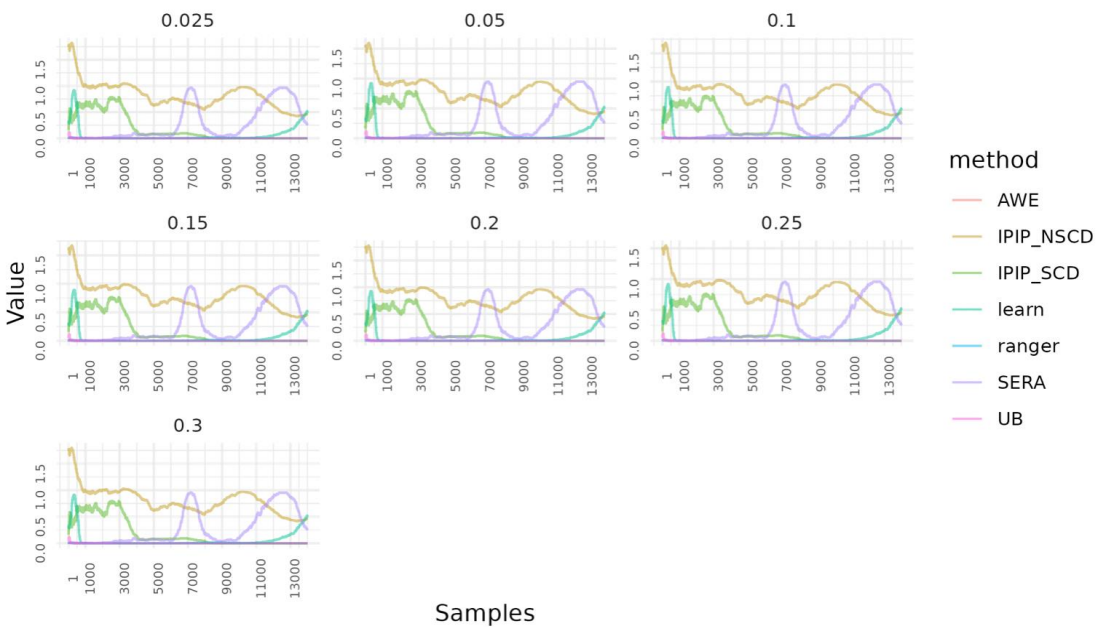
# Future work: Methods (in R)

- To the best of our knowledge, there is no R package that has passive models for training ensemble-based models to deal with concept drift in R.
- We have adapted the code in R for 8 of them.
- Developing an R package with these models and launching them for any chunk-based dataset.

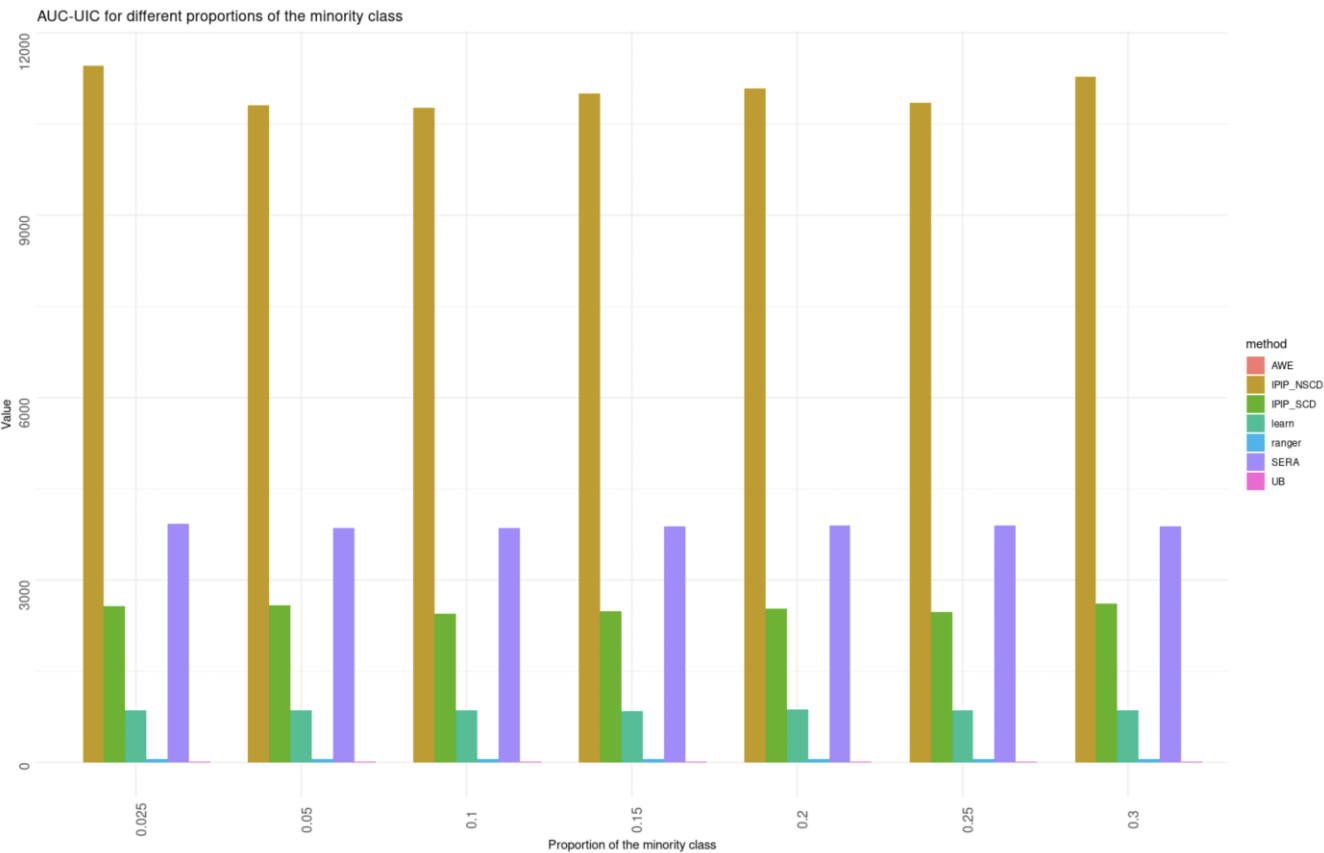


# Some Results in our COVID-19 dataset

UIC evolution for different values for the minority class proportion

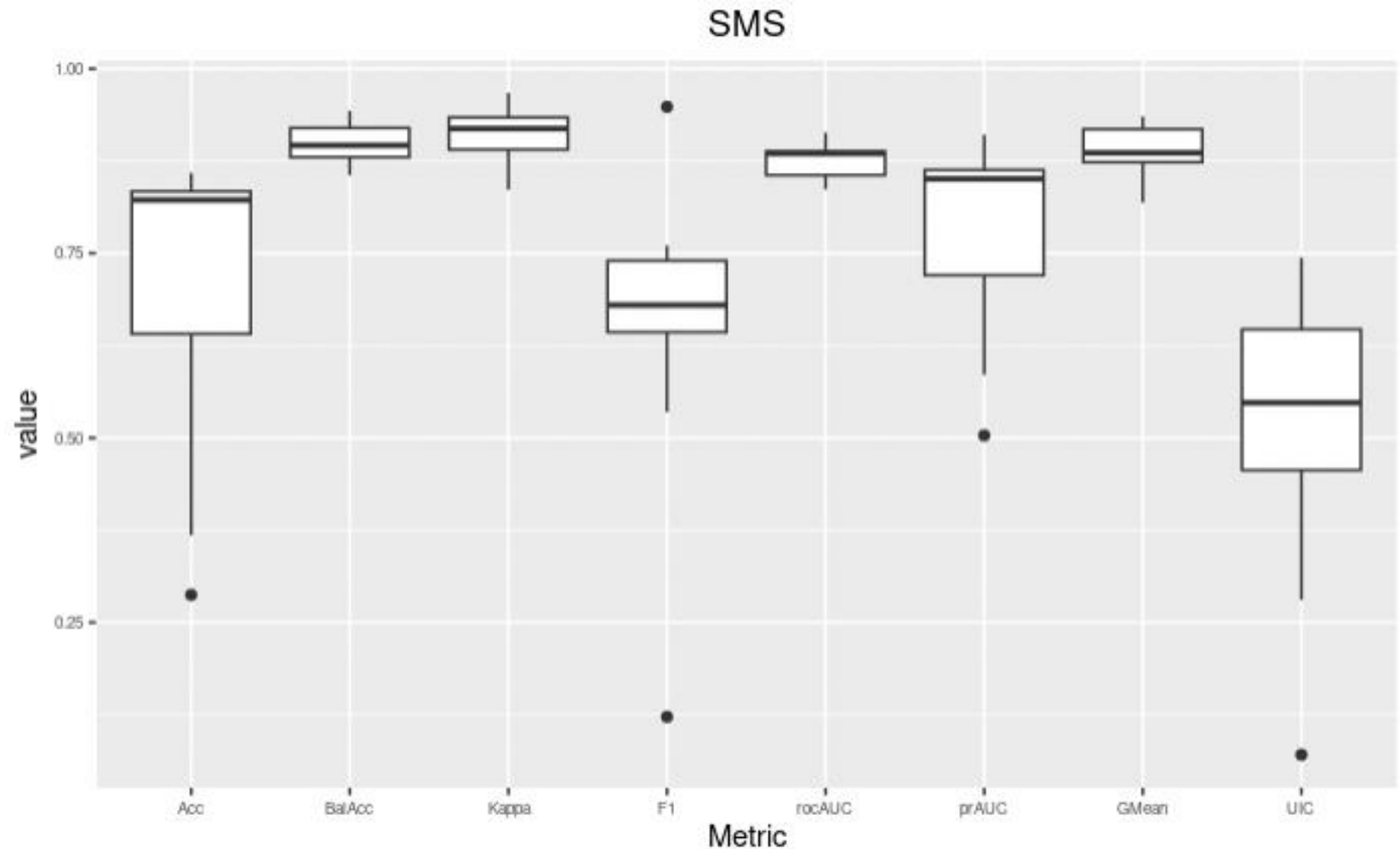


AUC



# Correlation between metrics and the IR for the COVID-19 dataset

- UIC is the less correlated measure with the imbalanced ratio compared to the rest of basic metrics for classification.
- Experiments in the remaining datasets are still in progress.



Thank you! :-)