CLASSIFIER EVALUATION OF AN IMBLANCED DATA SET OF THE PRESENCE OF FISH SPECIES FOR A STUDY DURING THE EARLY YEARS OF THE MIRGENBACH RESERVOIR.

- Samudranil BASAK[1], Gerard MASSON[2], Hoai Minh LE[1] and Baba Issa CAMARA[1]

[1] LCOMS, Universite de Lorraine, Site de Metz, 3 rue Augustin Fresnel, 57073, Metz, France

[2] LIEC, Universite de Lorraine, Site de Metz, Campus Bridoux, Rue Delestraint, 57070, Metz, France

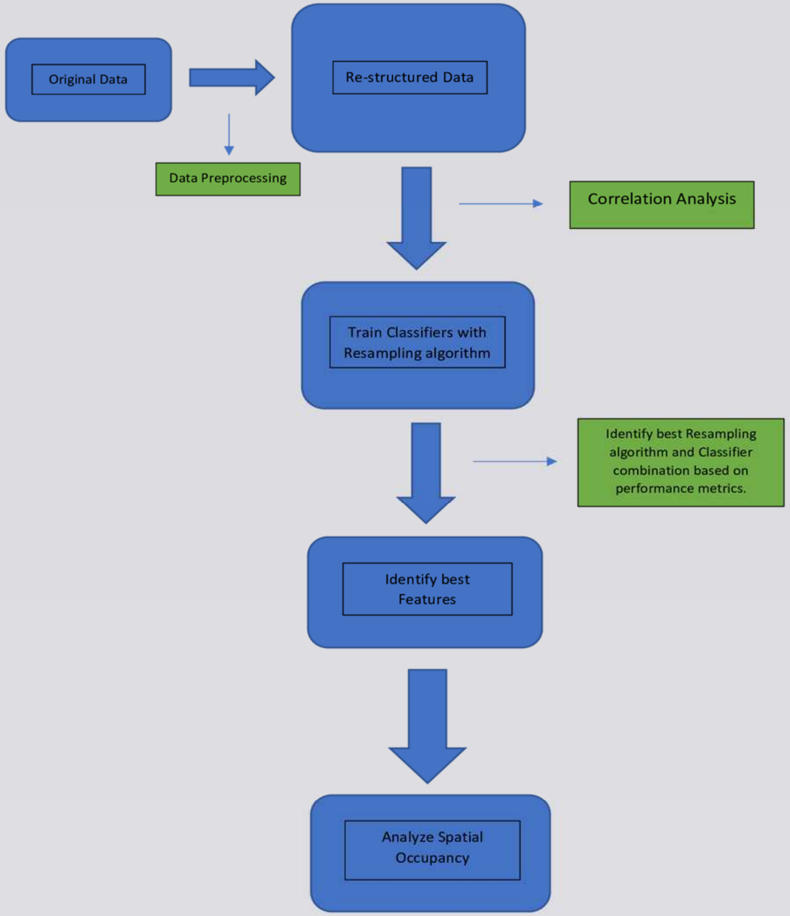
Introduction

- Mirgenbach reservoir was built in1985 as a supply of cooling water and buffer to the Cattenom Nuclear Electricity production center.
- The reservoir is characterized by relatively low species richness.
- The reservoir is characterized by relatively high water temperature and productivity. (Vein et al. 1990).
- There is an inlet of water from the Moselle river through the atmospheric cooling towers of the Cattenom Nuclear power station.
- The water after cooling down for some days (mean 15 days with 4 reactors working) is pumped out back to the Moselle.

Objective of study

- The initial study (done by A. Flesch, R. Marzou, G. Masson, P. Usseglio-Polatera and J.C. Moreteau) was to explain the variability in fish assemblage composition in the light of local environmental conditions.
- The main objective will be to analyze the Spatial occupancy of the fish species.
- To understand why these localizations were formed during the early days of the Mirgenbach.
- Which variables drove the presence and the fish species assemblages across multiple sites of the reservoir.

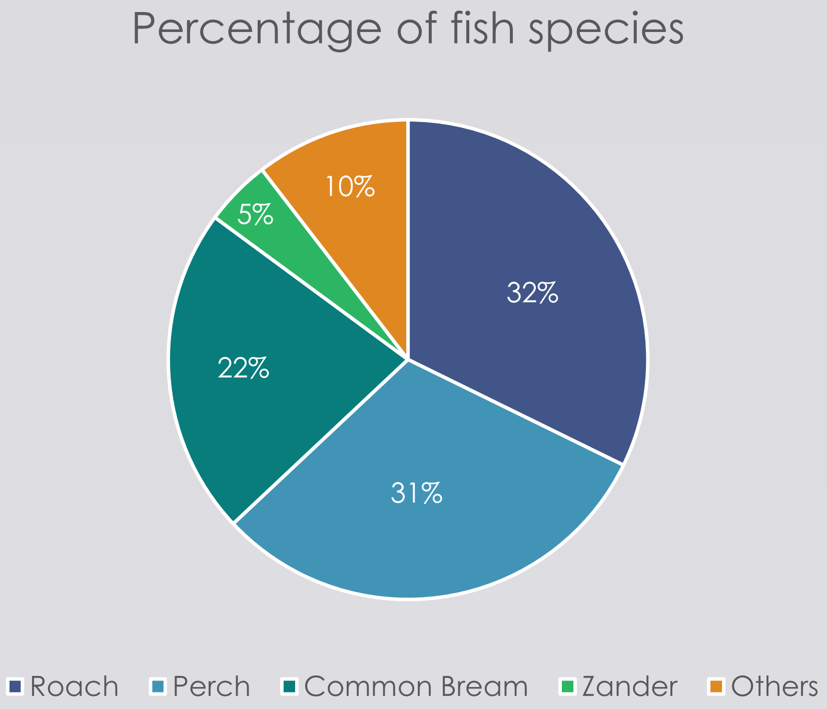
Study Design


Original Data
Re-structured Data
Data Preprocessing
Correlation Analysis
Train Classifiers with Resampling algorithm
Identify best Resampling algorithm and Classifier combination based on performance metrics.
Identify best Features
Analyze Spatial Occupancy

The Data set

- Surveys were conducted from October 1990 to October 1991 over 5 campaigns.
- During this period, 12 species of fish were caught.
- The survey covered the months of October 1990 (Autumn), November 1990(Winter), February 1991(Spring), June 1991(Summer) and October 1991 (Autumn).

Composition of Species and the Data set



- **Environmental variables -> 19**
- **Station Characteristics -> 14**
- **Temporal variables -> 3**
- **Biology of Fish -> 4**

Structure of the Mirgenbach reservoir.

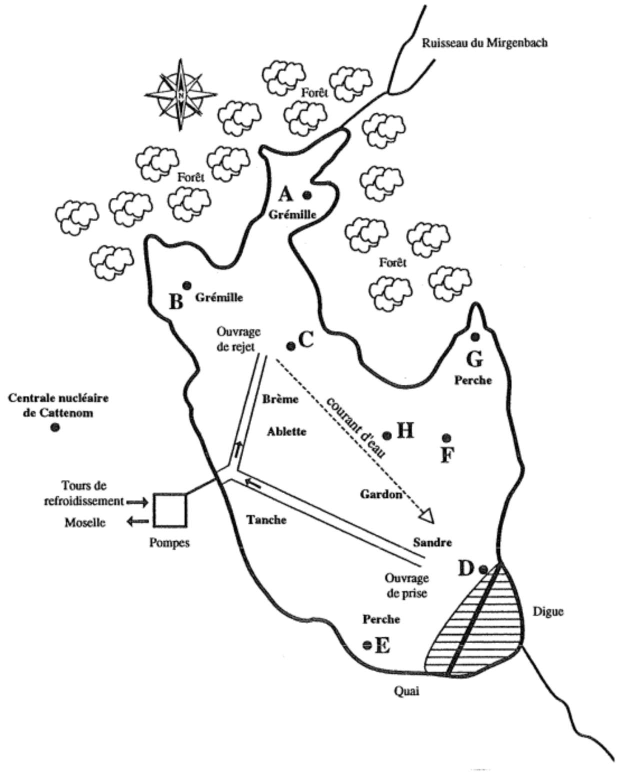


Figure 23: Typologie de la répartition des poissons dans le réservoir du Mirgenbach d'après l'AFCm.

Source: Fleshch A, 1994

- The reservoir has 8 sampling stations.
- Stations A, B, C and G are quite near to a forest shoreline, thus there is vegetation.
- Stations D and E are near the docks.
- There is a source of water inlet and water outlet to and from the reservoir.
- Station C is located near the water inlet and D is located near the water outlet.

Characteristics of each Station

variables	codes	modalities or classes	qualification station	A	B	C	D	E	F	G	H
water depth	dep 1	<= 5m	ok	dep 1	dep 1			dep 1		dep 1	
	dep 2	> 6-12m	ok			dep 2					dep 2
	dep 3	> 13-20m	ok				dep 3		dep 3		
nearest shore type	typ 1	naked bank	ok					typ 1			
	typ 2	blocks	ok				typ 2		typ 2		
	typ 3	vegetation	ok	typ 3	typ 3	typ 3				typ 3	typ 3
distance from water inlet	dou 1	<= 100m	ok			dou 1					
	dou 2	> 100-350m	ok	dou 2	dou 2						dou 2
	dou 3	> 350m	ok				dou 3	dou 3	dou 3	dou 3	
distance from water output	din 1	<= 250m	ok				din 1	din 1	din 1		din 1
	din 2	> 250-450m	ok			din 2				din 2	
	din 3	> 450m	ok	din 3	din 3						
distance to nearest shore	dis 1	<= 30m	ok		dis 1		dis 1	dis 1		dis 1	
	dis 2	> 30-80m	ok	dis 2		dis 2					
	dis 3	> 80m	ok						dis 3		dis 3
granulometry of the nearest littoral substrate	sub 1	sludge	ok	sub 1	sub 1	sub 1				sub 1	sub 1
	sub 2	pebbles	ok					sub 2			
	sub 3	blocks	ok				sub 3		sub 3		
coves	cov 1	no coves	ok			cov 1	cov 1	cov 1	cov 1		cov 1
	cov 2	upstream from water inlet	ok	cov 2	cov 2						
	cov 3	downstream from water output	ok							cov 3	
substrate heterogeneity	stu 1	low	ok				stu 1	stu 1	stu 1		stu 1
	stu 2	middle	ok			stu 2				stu 2	
	stu 3	high	ok	stu 3	stu 3						
wind exposure	win 1	low	ok	win 1						win 1	
	win 2	middle	ok		win 2						
	win 3	high	ok			win 3	win 3	win 3	win 3		win 3
trophic potential	tro 1	low	ok	tro 1	tro 1			tro 1			
	tro 2	middle	ok			tro 2					tro 2
	tro 3	high	ok				tro 3		tro 3	tro 3	
tree stumps	smp1	close (<30m)	ok	smp1	smp1					smp1	
	smp2	middle (30 à 100m)	ok			smp2					smp2
	smp3	distant (>100m)	ok				smp3	smp3	smp3		

Algorithms and Models considered

Resampling algorithms:

Under-sampling:

- Condensed Nearest Neighbor

Over-sampling:

- Adaptive synthetic sampling (ADASYN)
- Synthetic Minority Oversampling (SMOTE)
- Borderline SMOTE

Combined Methods:

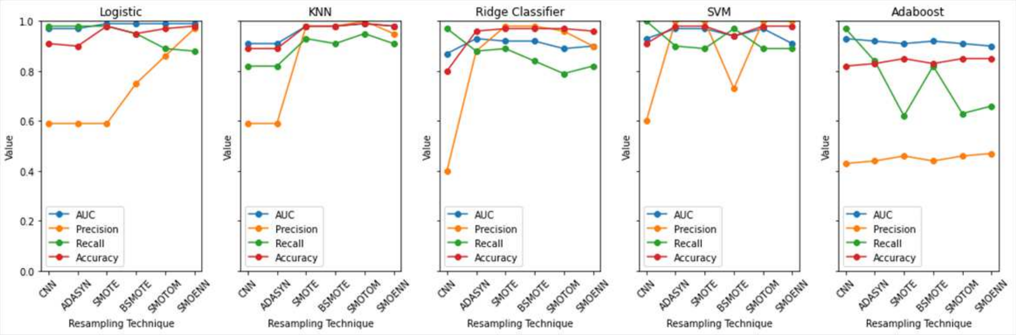
- SMOTE + Tomek Link Removal
- SMOTE + Edited Nearest Neighbors

Classifiers:

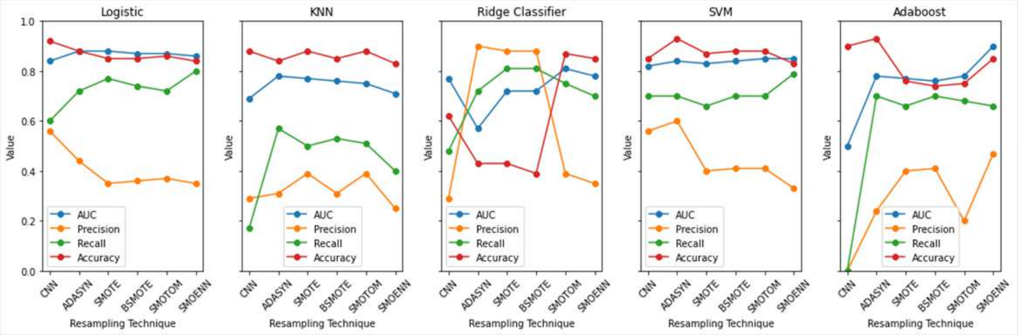
- Logistic
- Kth Nearest Neighbors
- Ridge Classifier
- Support Vector Classifier (SVC)
- Adaptive Boosting (with a Decision stump as the base estimator)

Results

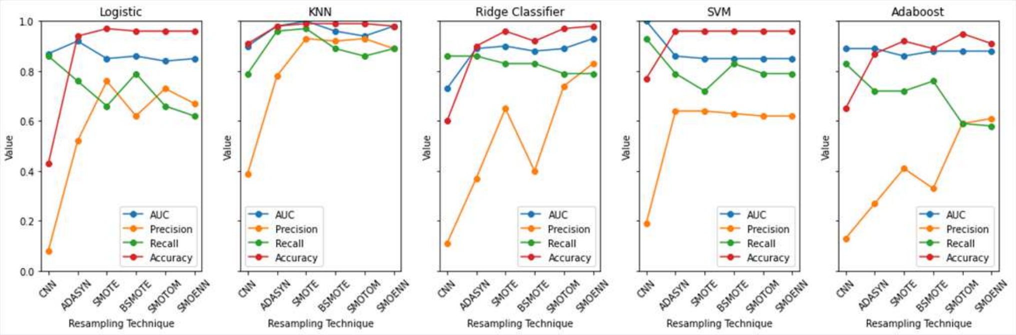
Metric evaluation per Classifier



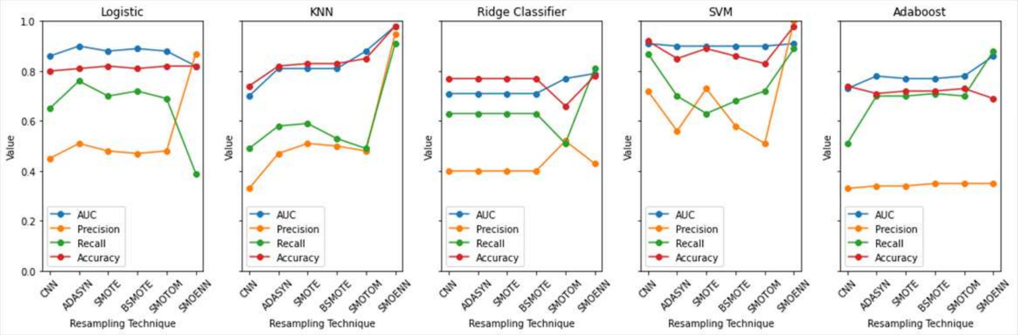
Bream



Other

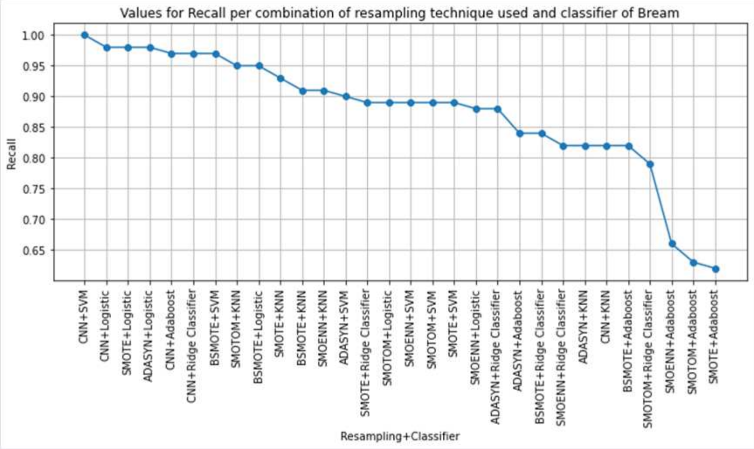


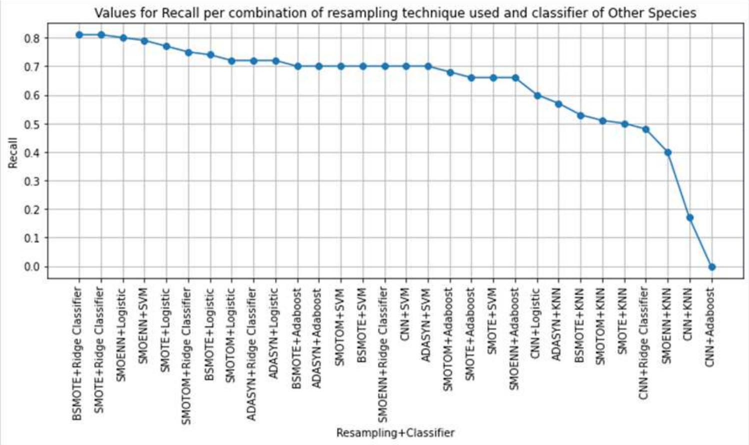
Perch

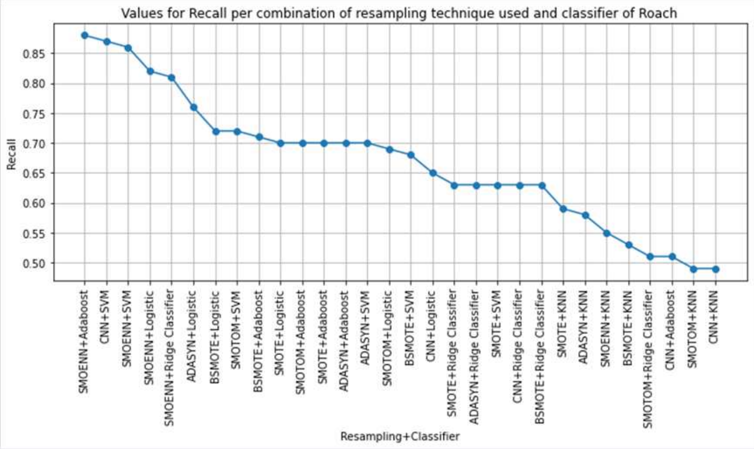


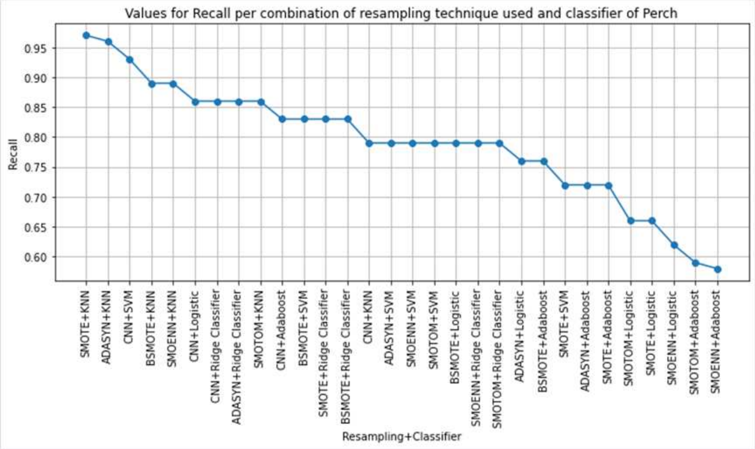
Roach

Results

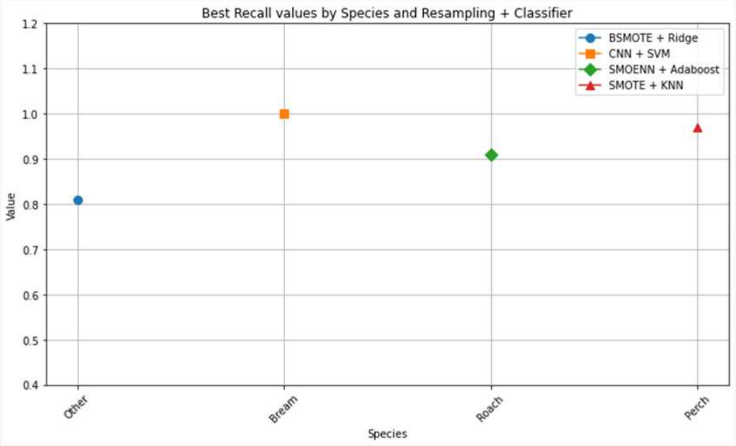

Values for Recall per combination of resampling technique used and classifier of Bream

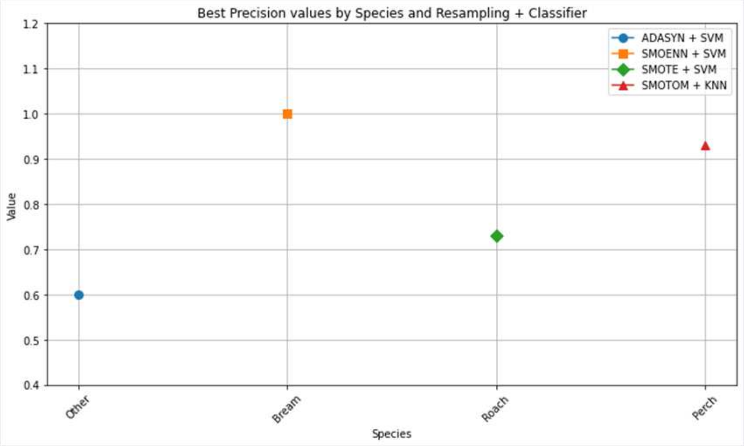

Values for Recall per combination of resampling technique used and classifier of Other Species

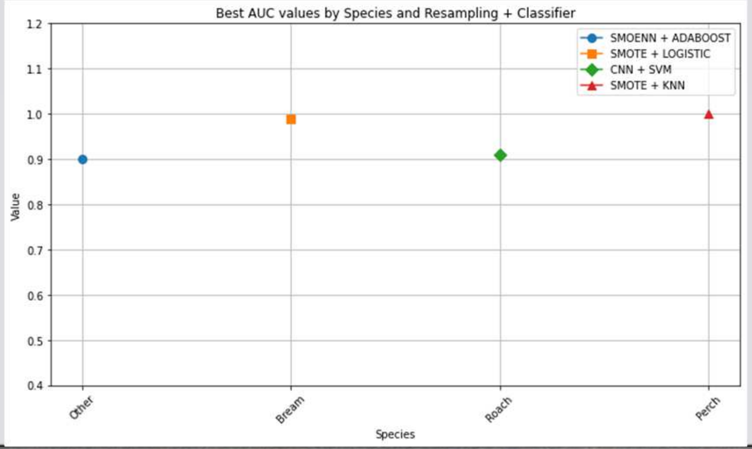

Values for Recall per combination of resampling technique used and classifier of Roach


Values for Recall per combination of resampling technique used and classifier of Perch

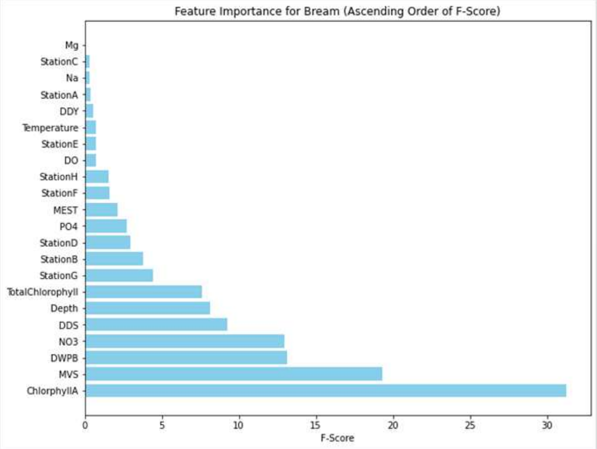
Best Resampling and Classifier combination by best metric

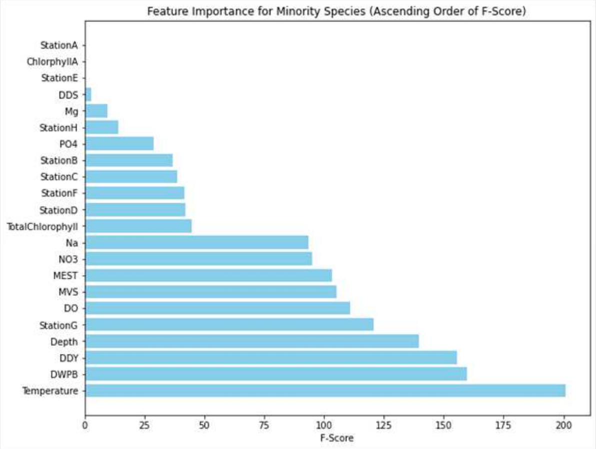


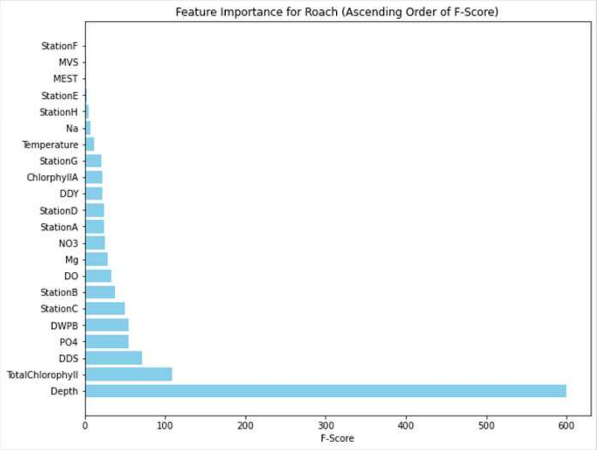


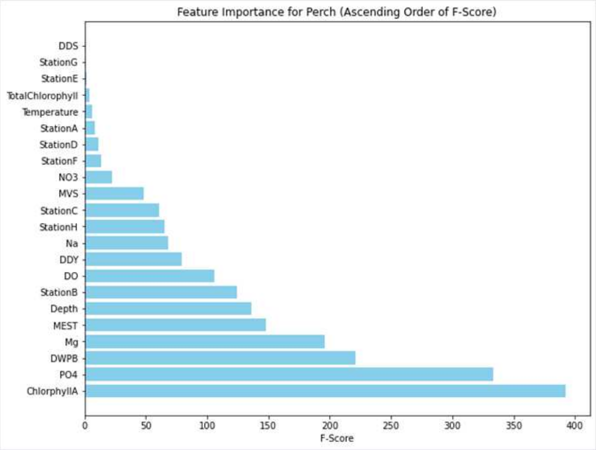


Feature Importance



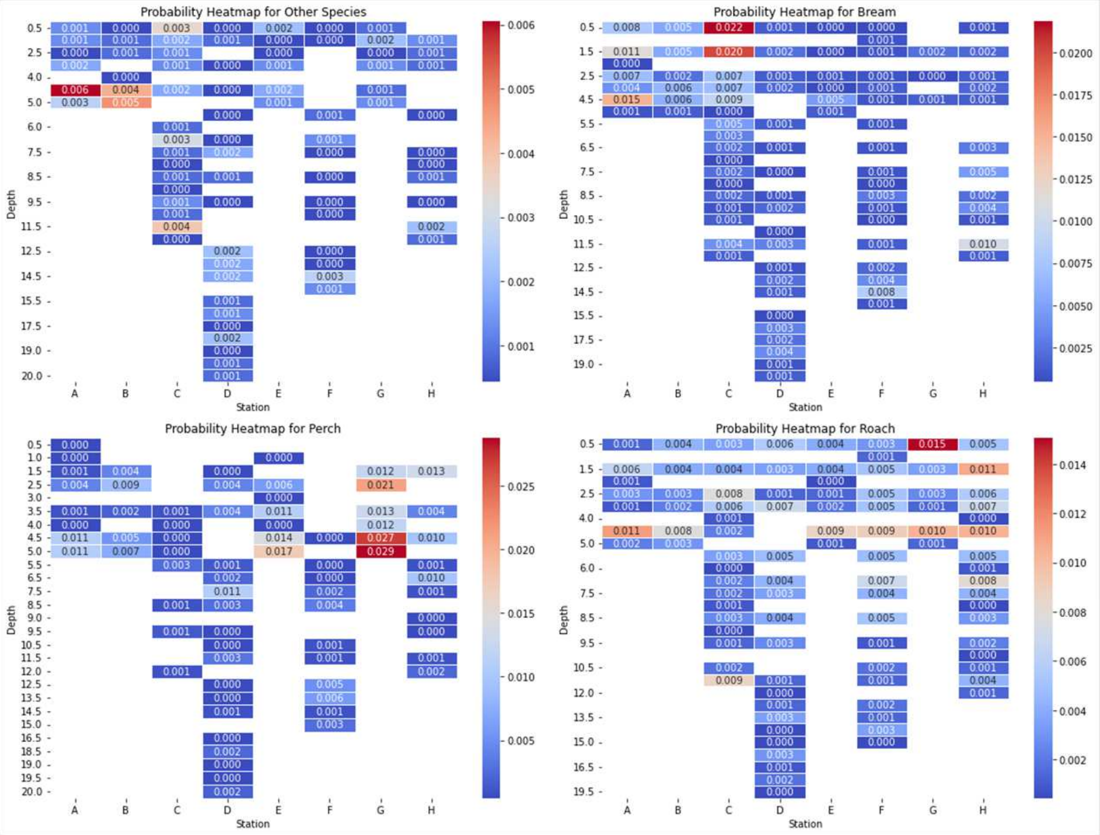






- **List of important variables driving presence of each species:**
- For Bream: Active Chlorophyll, Volatile Suspended Matters, Dry weight of Phytoplankton Biomass, Nitrate Ion concentration, Cumulative Degree Days of the season, Depth of Station and Total Chlorophyll.
- For Roach: A lot of the same, except Station Depth is a critical factor.
- For Perch, Total Suspended Solids (MEST), Magnesium content and the Phosphate Ion concentration.
- For the minority species: Temperature is crucial. Dissolved Oxygen and Sodium concentration plays a lesser important role.

Spatial Occupancy of capturing species by Station and Depth



Discussion

- In this study, we first address the problem, an imbalanced data set poses to traditional classifiers. 6 resampling methods are used with 5 classifiers and their performance is evaluated. We show the best classifier and re-sampler combination with the respective metric.
- We then analyze the features that are important in driving the occupancy of each fish species of the Mirgenbach.
- Bream occupies mostly at Station C at shallow depths and at a depth of 4.5 meters in Station A.
- Perch and Roach tend to gather around Station G, with Roaches gathering mostly at a shallow depth of 0.5 meters and Perch mostly at 4.5 to 5 meters.
- Both Stations have a few things in common, like vegetation shore type, granulometry (sludge), absence of coves and moderate substrate heterogeneity.
- Vegetation shore type tends to give higher levels of Chlorophyll (active or Inactive) and we see in both Bream and Perch, Chlorophyll content is of significant importance.
- Most of the members of the Minority species have occupancy in Stations A and B at depths of between 4 and 5 meters with a small percentage occupying lower depths of 11.5 meters.

References

1. Gulnaz Ahmed, Meng Joo Er, Mian Muhammad Sadiq Fareed, Shahid Zikria, Saqib Mahmood, Jiao He, Muhammad Asad, Syeda Fizzah Jilani, and Muhammad Aslam. Dad-net: Classifica- tion of alzheimer's disease using adasyn oversampling technique and optimized neural network.*Molecules*, 27(20):7085, 2022.
2. Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behaviorof several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.
3. Jakob Brandt and Emil Lanzén. A comparative review of smote and adasyn in imbalanced data classification. 2021.
4. Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM SIGKDD explorations newsletter*, 6(1):1–6, 2004.
5. Samuel Dembski. *Stratégies d'occupation spatiale en milieu lacustre: réponses de l'ichtyofaune dans un réservoir échauffé, non stratiFIé*. PhD thesis, Université de Metz, 2005.
6. A Flesch. Biologie de la perche (perca fluviatilis) dans le réservoir du mirgenbach (cattenom, moselle)[phd thesis]. *Université de Metz, France*, page 241, 1994.
7. Anne Flesch, Gérard Masson, and Jean-Claude Moreteau. Temporal distribution of perch (perca fluviatilis l.) in a lake-reservoir (moselle, france): analysis of catches with vertical gill nets. *Hydrobiologia*, 300:335–343, 1995.
8. Vaishali Ganganwar. An overview of classification algorithms for imbalanced datasets. *Inter- national Journal of Emerging Technology and Advanced Engineering*, 2(4):42–47, 2012.
9. Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.
10. Peter Hart. The condensed nearest neighbor rule (corresp.). *IEEE transactions on information theory*, 14(3):515–516, 1968.

Thank you!