AIMS African Institute for Mathematical Sciences SENEGAL

Project 1: Ensemble Learning

Presented by:

ADOU Kouamé Mathurin

KENNE Annine Duclaire

Nancy Fosua Boakye

MEGNE CHOUDJA Pauline Ornela

Introduction

Models

Data Preprocessing

Results and Discussion

Conclusion

Introduction

To help create innovative model which better improves ML models, we carried out this project.

Objectives

- **Analyze and preprocess the data.**
- **Build and train different models including logistic regression, SVM, FCN and decision tree.**
- **Combine those models based on different ensemble learning approaches.**

Machine Learning Models

Ensembling Learning is a meta approach which aims to seek better performance by combining the predictions from multiple models.

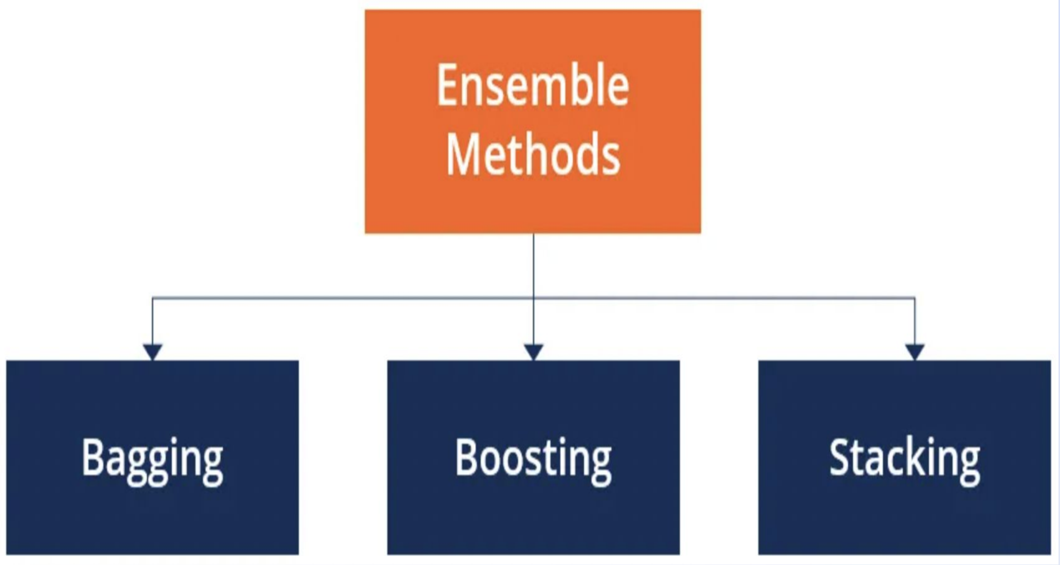


Figure 1 : Ensemble methods

a. Bagging Method

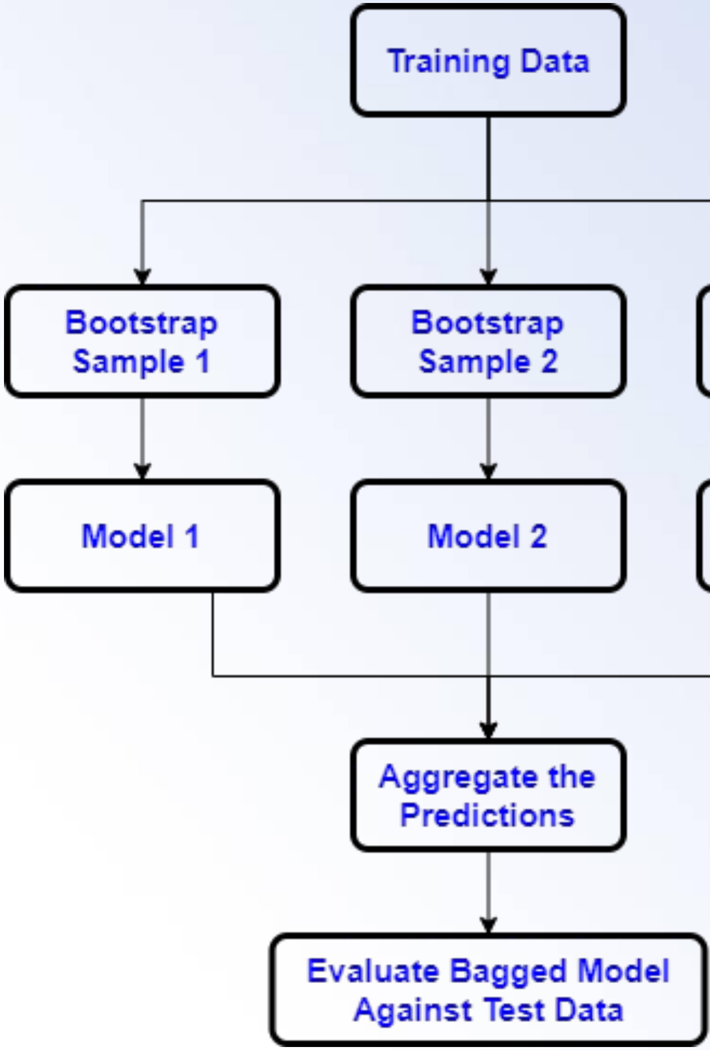


Figure 3 : Bagging principle

b. Boosting Method(Adaboost)

step 1: Assign equal weights to each point in the dataset,

$$w_i = \frac{1}{N}, \, i = 1, \ldots, N \tag{1}$$

step 2: Create the first base learner.

step 3: Compute the total error,

$$TE = \sum_{i=1}^{N} (y_i \neq \hat{y_i}) w_i \tag{2}$$

step 4: Compute the performance of the model,

$$P = \frac{1}{2} log\left(\frac{1 - TE}{TE + \epsilon} + \epsilon\right) \tag{3}$$

Machine Learning Models

step 5: Updating weight,

$$\begin{cases} new\_weight_{(y_i \neq \hat{y}_i)} = old\_weight_{\,(y_i \neq \hat{y}_i)} \cdot e^{P} & (a) \\ new\_weight_{(y_i = \hat{y}_i)} = old\_weight_{\,(y_i = \hat{y}_i)} \cdot e^{-P} & (b) \end{cases} \tag{4}$$

step 6:Randomly sample a new dataset.

step 7: Build a new model.

step 8: Repeat the same process from step 2 until the total number of fixed models.

step 9: Make predictions using test data.

Machine Learning Models

c. Voting Method

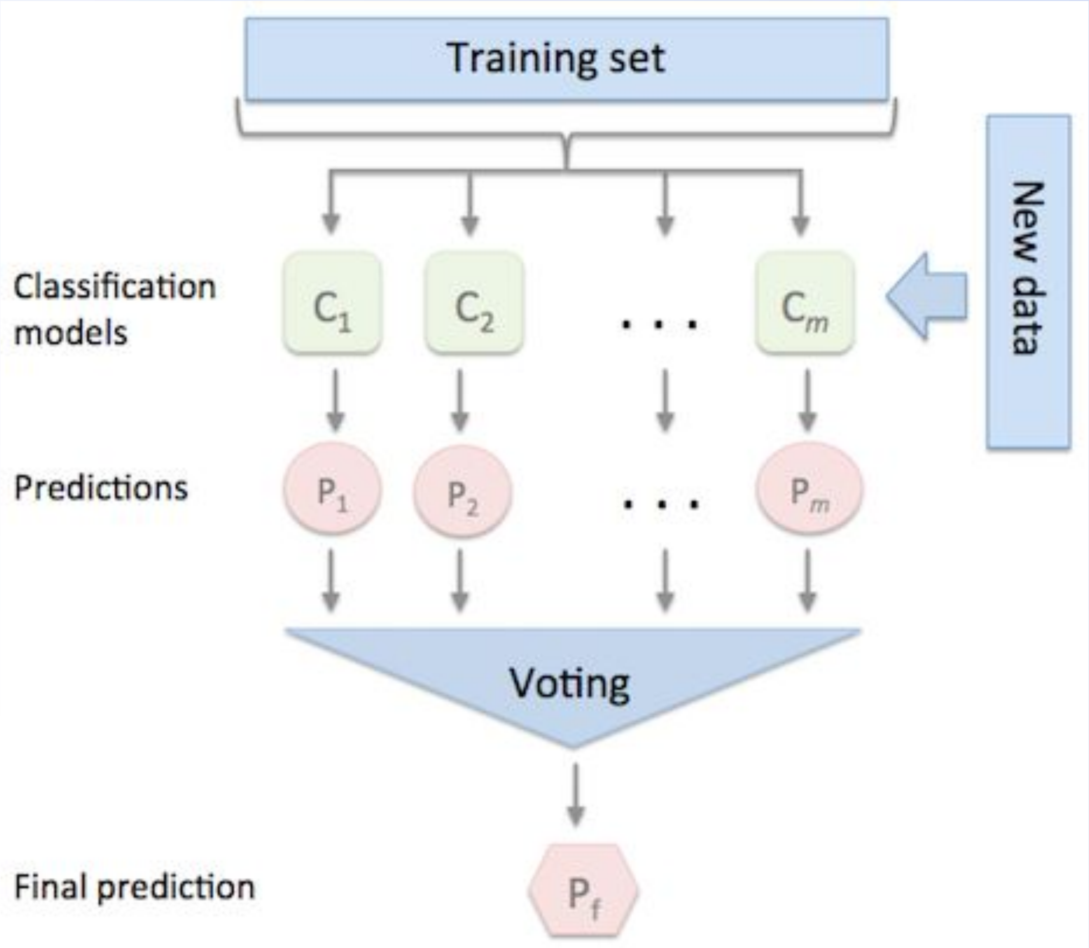


Figure 4: Voting principle

Data Preprocessing

- Convert object feature to numerical values using one hot and label encoding.
- Oversampling and undersampling techniques to solve the issue of imbalanced data.

Figure 5: Class representation

- Random Oversampler
- Random Undersampling

Results and Discussions

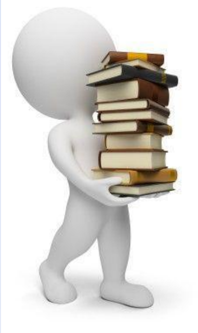
Single Models				
Models		Precision 2	Precision 4	Accuracy
Logistic		1.00	0.95	0.987
SVM		1.00	0.95	0.987
DT		0.99	0.88	0.962
FCN		1.00	0.93	0.981
Ensemble Models				
Voting		1.00	0.95	0.987
Bagging	Logistic	1.00	0.95	0.987
	SVM	1.00	0.95	0.987
	DT	1.00	0.93	0.981
	FCN	1.00	0.95	0.987
Boosting	Logistic	1.00	0.88	0.987
	SVM	1.00	0.88	0.987
	DT	1.00	0.88	0.785
	FCN	1.00	0.88	0.968

Figure 6: Models summary

Conclusion

Based on our findings, we can conclude that:

- ❖ **Label encoding performs better than the one hot encoding.**
- ❖ **The oversampling approach performs better than the undersampling.**
- ❖ **Having a good accuracy doesn’t always implies good model.**
- ❖ **Ensembles methods improves the performance of a model, however, it is not guaranteed for all cases.**



THANK YOU FOR YOUR ATTENTION