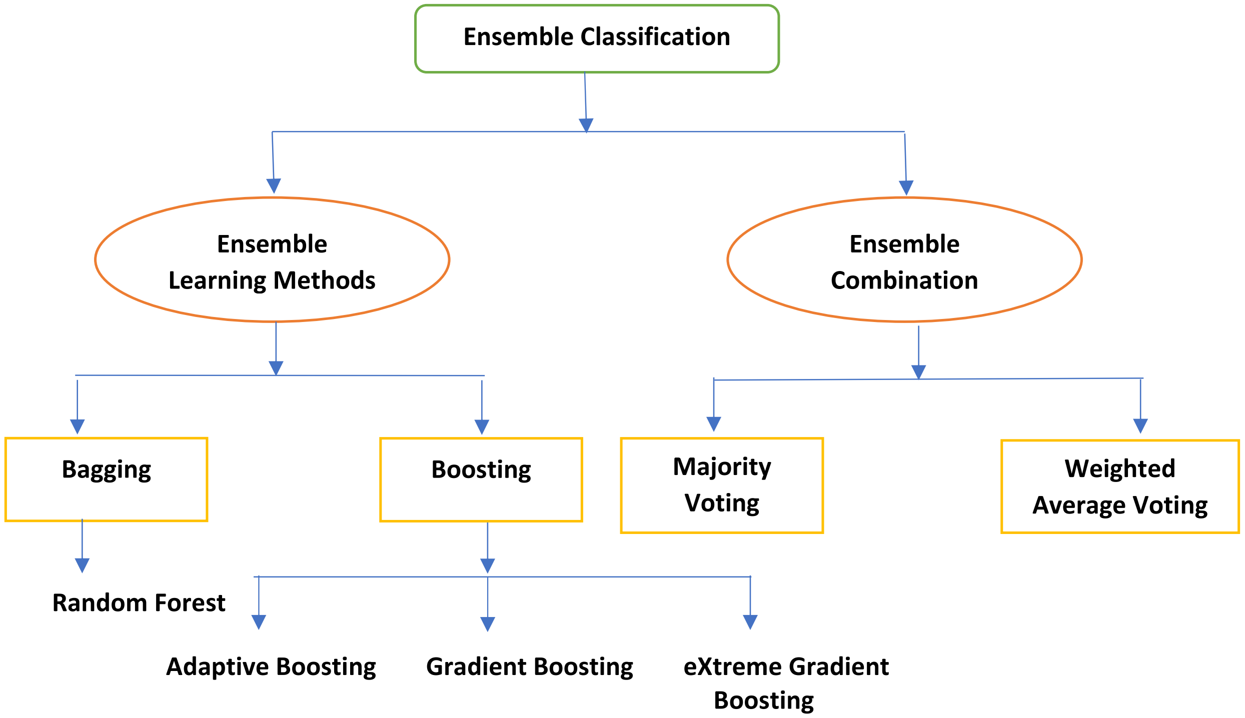
CREDIT CARD FRAUD DETECTION WITH BAGGING ENSEMBLE LEARNING



Ensemble refers to the combination of several distinct set of individual machine learning techniques together in order to enhance the strength, efficiency and predictive performance of the model.

Ensemble Learning Methods

Ensemble learning methods can be divided into two groups: *Sequential ensemble methods*, where the base learners are generated sequentially and *Parallel ensemble methods*, where the base learners are generated in parallel. The key objective of the ensemble learning is to reduce bias and variance.

Bagging

Bagging (Bootstrap aggregation) methods form a class of algorithms which build several instances of a black-box estimator on random subsets of the original training set and then aggregate their individual predictions to form a final prediction. On average, the combined estimator is usually better than any of the single base estimator because its variance is reduced.

Boosting

In Boosting methods, base estimators are built sequentially and one tries to reduce the bias of the combined estimator. The motivation is to combine several weak models to produce a powerful ensemble.

Ensemble Combination Rules

The machine learning classifiers can be combined using different combination rules i.e., Majority Voting and Weighted Average Voting.

Majority Voting

Every individual model makes a prediction for each test instances and the final output prediction is the one that receives the majority of votes.

Weighted Average Voting

In weighted average voting we can increase the importance of one or more models by assigning weights to the models. The prediction of each model is multiplied by the weight and then their average is calculated.

BAGGING WITH PYTHON

Data Source: https://www.kaggle.com/janiobachmann/credit-fraud-dealing-with-imbalanced-datasets/data

Context

It is important that credit card companies are able to recognize fraudulent credit card transactions so that customers are not charged for items that they did not purchase.

Content

The dataset contains transactions made by credit cards in September 2013 by European card-holders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'.

Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependent cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.