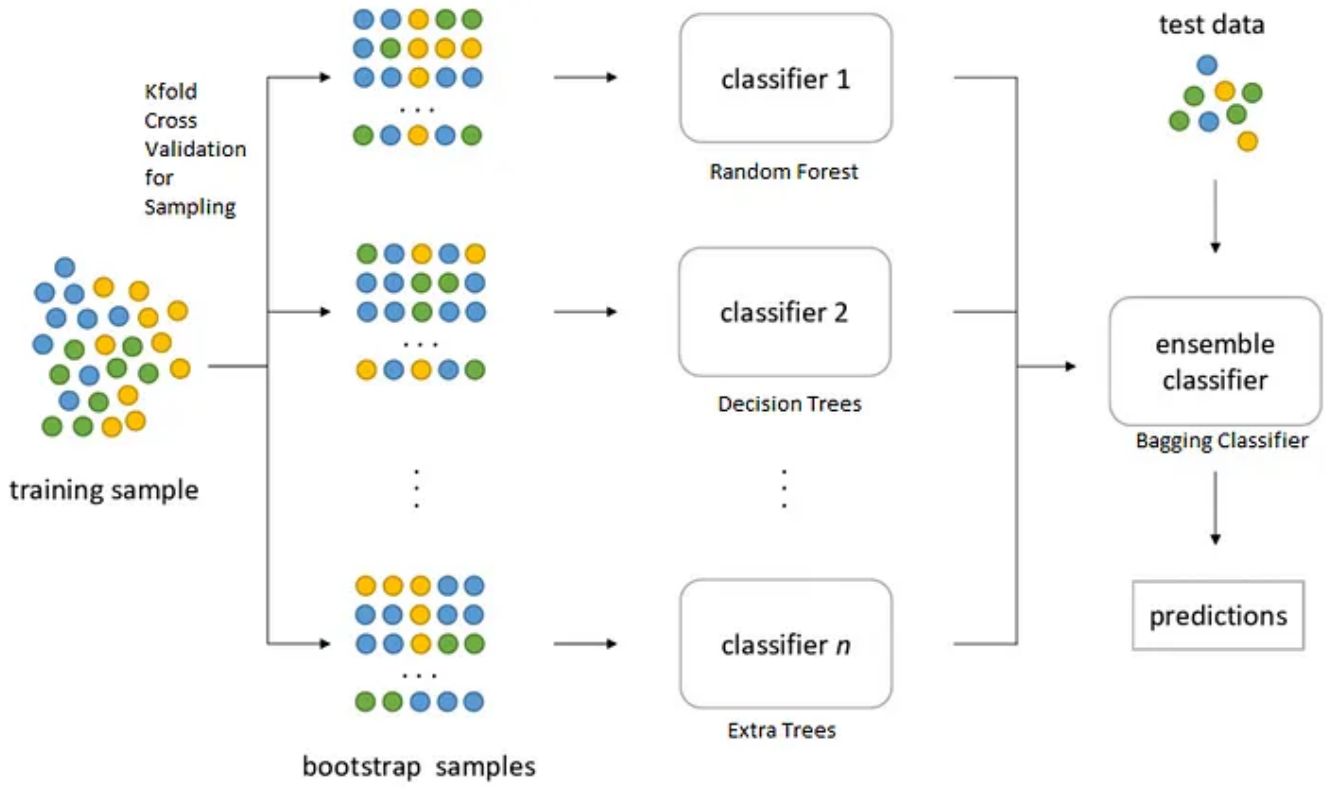
Bagging — Ensemble meta Algorithm for Reducing variance

Ensemble learning Series...!!!



Bagging Classifier Process Flow

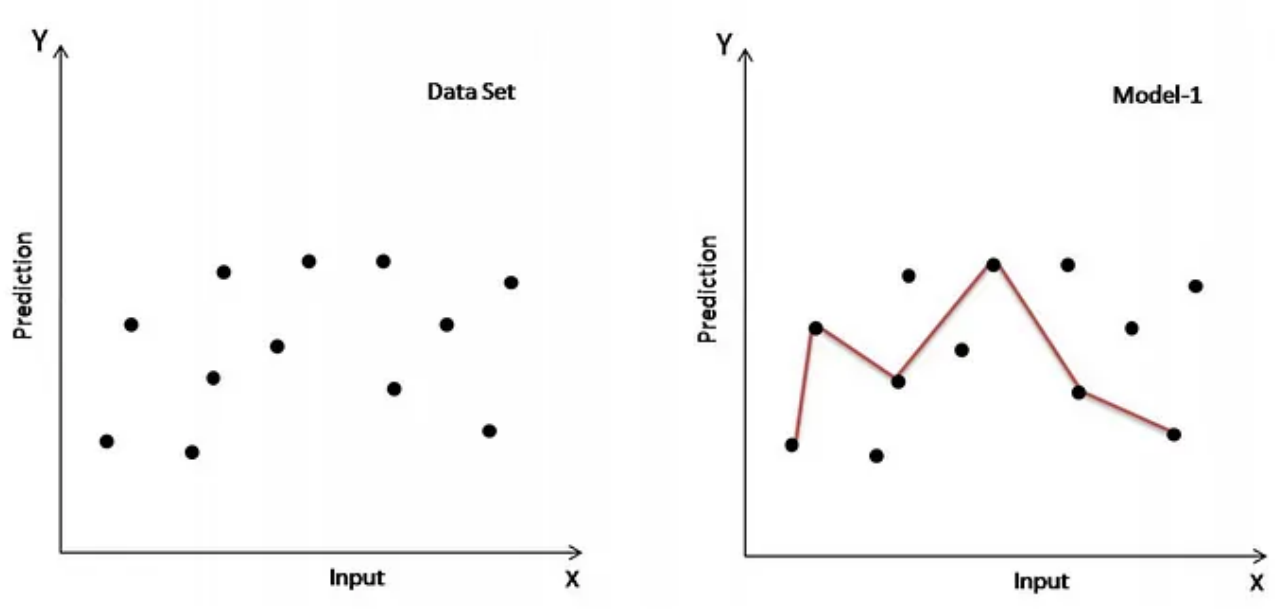
Ensemble learning Series — (Happy Learning...!!!)

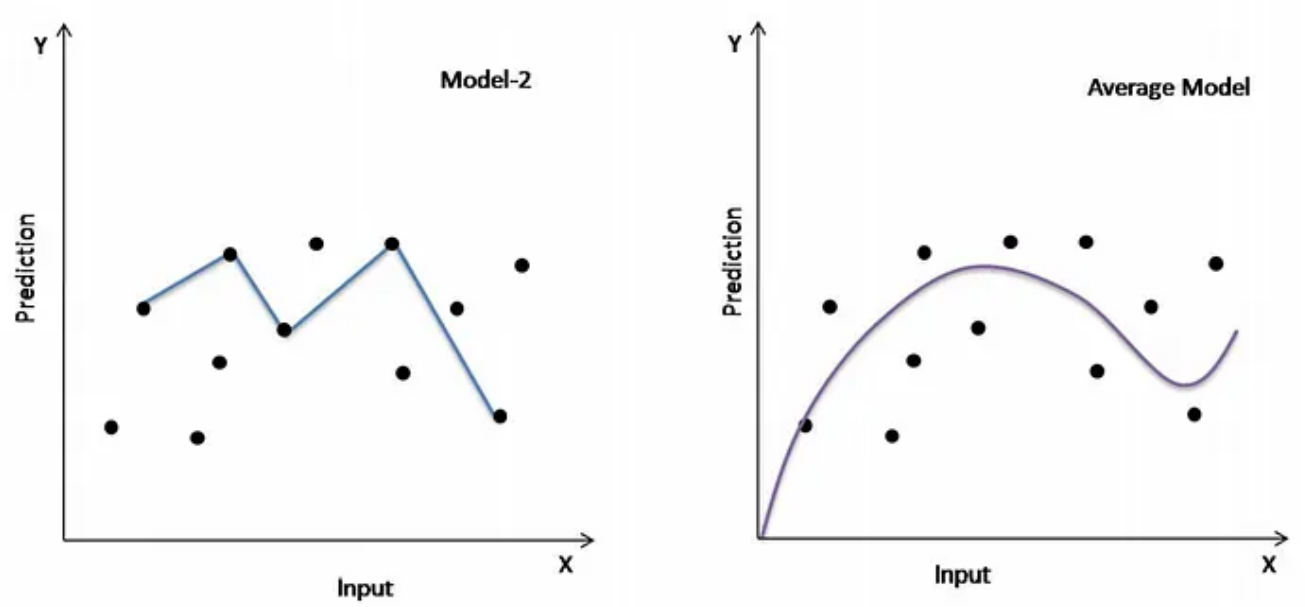
1. Ensemble Learning Relation With Bias and variance
2. Ensemble Learning- The heart of Machine learning
3. Bagging — Ensemble meta Algorithm for *Reducing variance*
4. Boosting- Ensemble meta Algorithm for *Reducing bias*
5. Stacking -Ensemble meta Algorithms for *improve predictions*
6. Ensemble learning impact on Deep learning

Bagging — The Enemy of the Variance

Bagging, an acronym for ***bootstrap aggregation***, creates and replaces samples from the data-set. In other words, each selected instance can be repeated several times in the same sample. We seem to increase our training data with bootstraps, which are each created and then used to create a classifier model. The final prediction is the average of all predictive models.

The most popular bagging algorithm commonly used by data scientist is the random forest based on the decision tree algorithm. Another useful algorithm is the pocket filling of the neighboring subspace closest to K (KNN), where basic students are based on the closest neighbor algorithm to k. We will discuss these algorithms in detail in the future. We can understand the crisis with the following example. Suppose we want to customize a complex data set, as shown in the following figure:

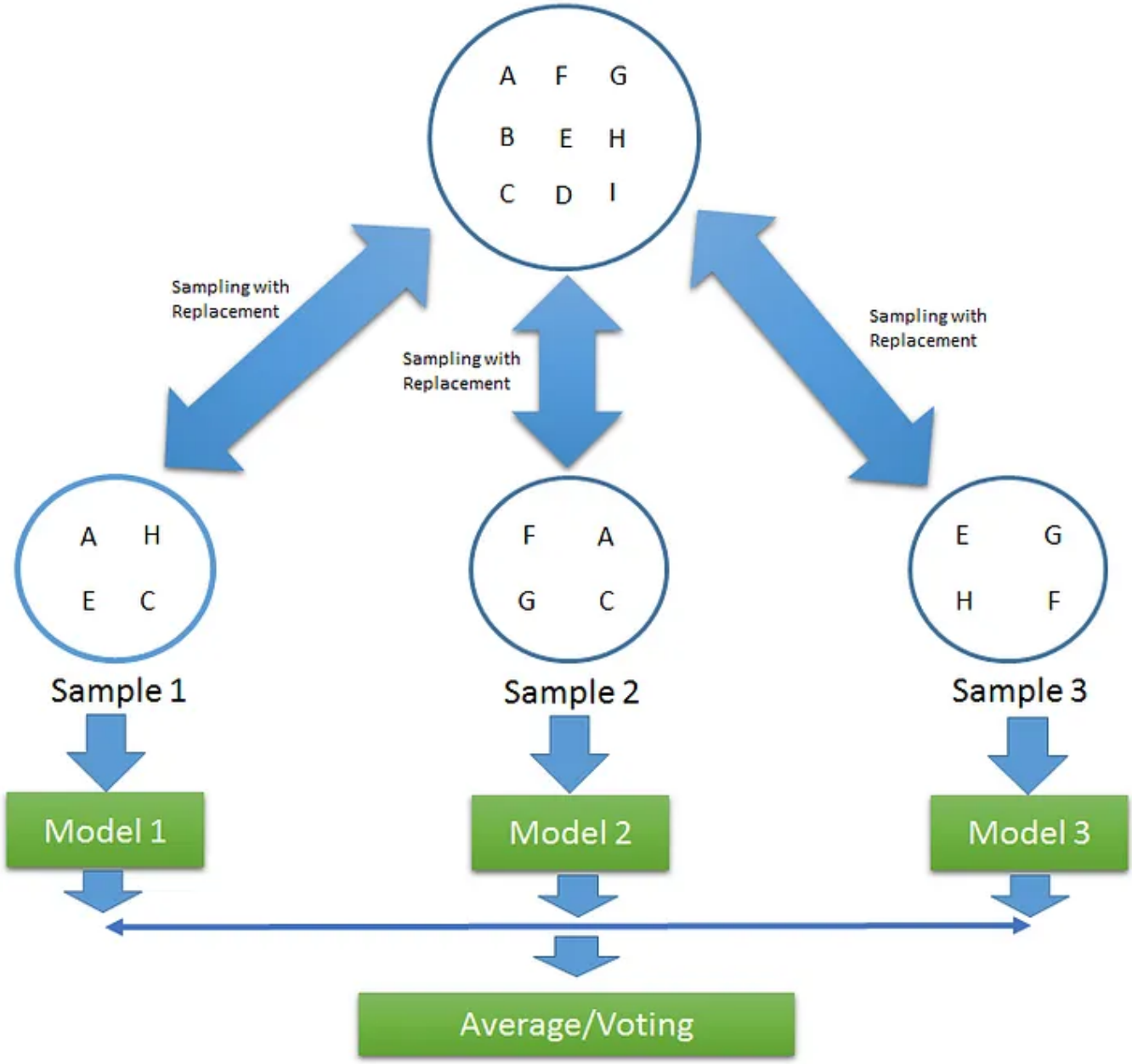




Four different plot shown in the figure.

- **Plot 1 :** Shows the Distribution of data. X axis represent input and y axis represent the output.
- **Plot 2 :** Once you want to customize your simple model using this data set. You can saw that our model is inefficient to fit on all data points or has a significant bias error, but some data points are well predicted by our model.
- **Plot 3:** Imagine that we can train Second model known as Model 2 to adjust our data still model is not performing well.However still this model 2 is performing well as compare to model 1.
- **Plot 4:** If we can average forecast for both models, we can get optimal result which is fit the data well with less bias than individual high bias.

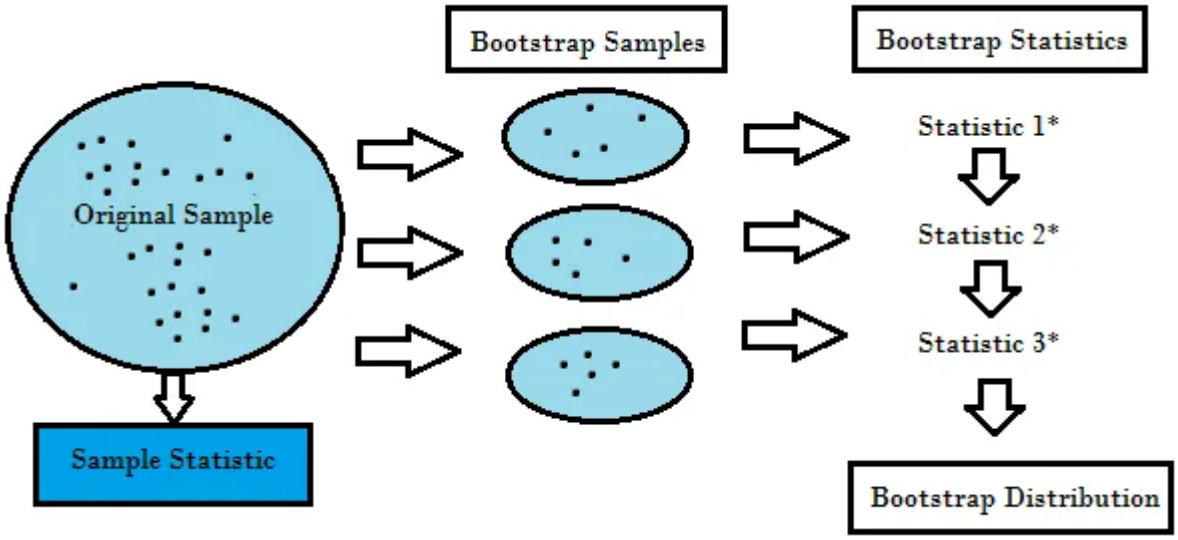
The above example show the power of ensemble which can combine multiple result of models below average and create combine model that can predict the optimal result with very high confidence.

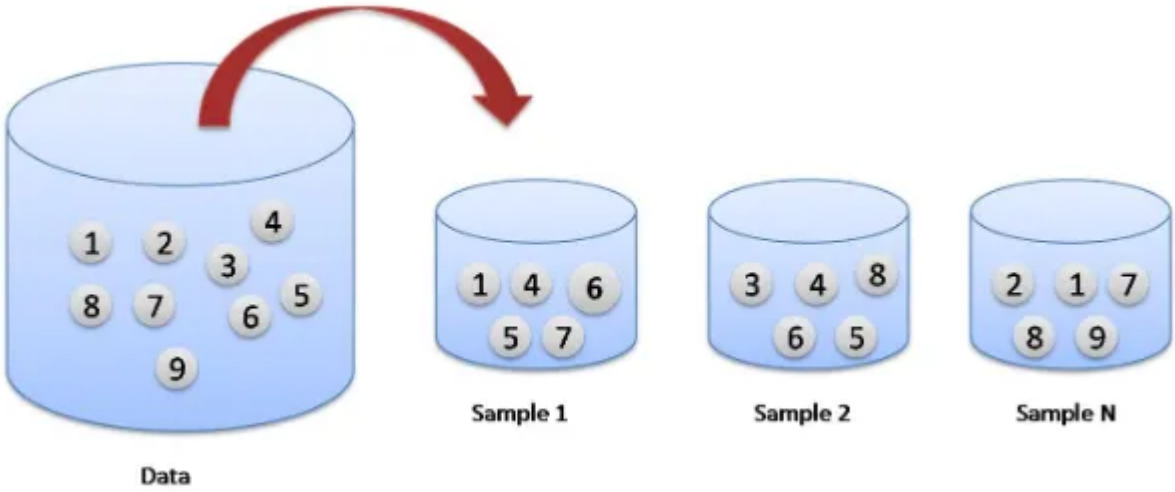


This image shows that the bagging example has three steps :

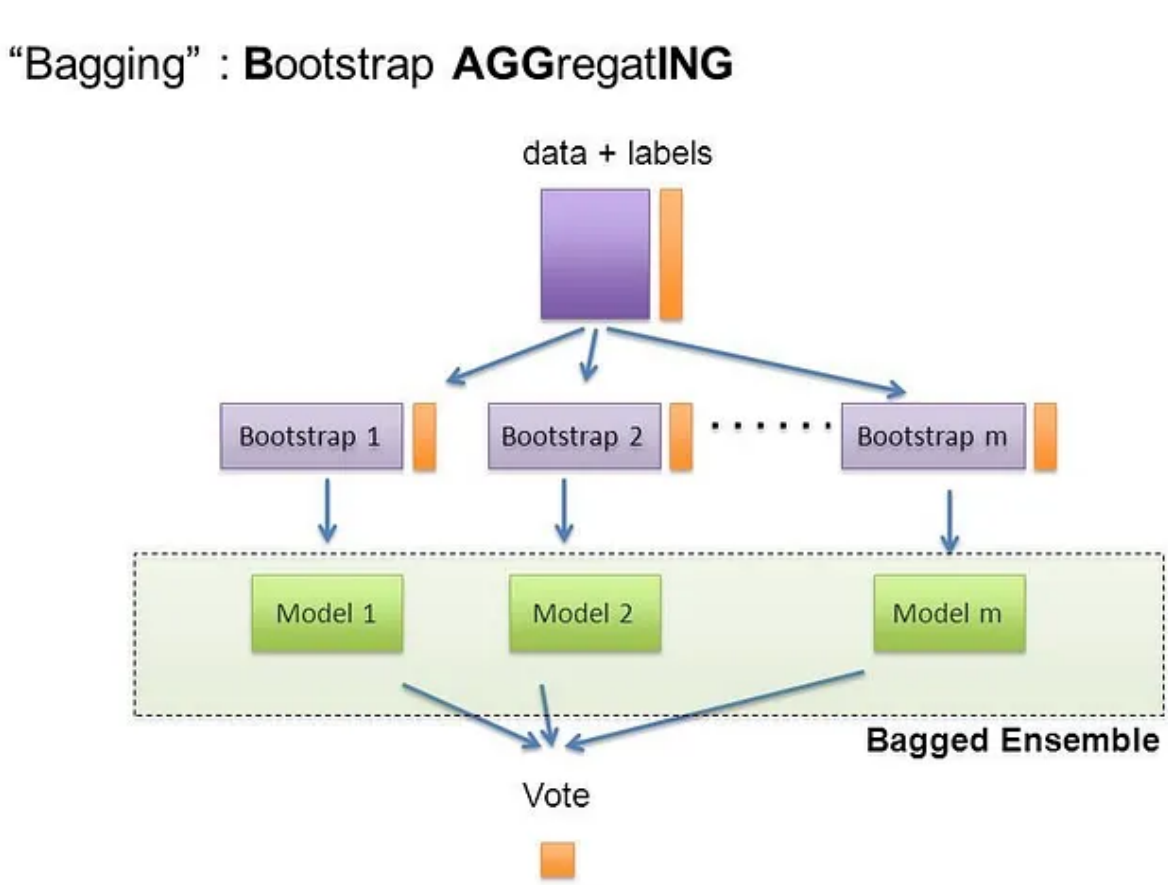
1. **Bootstrap data(Sampling)**
2. **Aggregation or Model fit**
3. **Combination different model with Result aggregation.**

Bootstrapping is a sampling technique in which we create multiple random sample from our training data-set. Below example can be created for selecting random data point instances and combine them which is shown in the below figure.





- In the above example, Imagine that we have one data bucket and we can randomly distribute the sample to different bucket in **sample 1, sample 2 and sample 3. if we can find like any instance is repeating** . That is, when we choose a new instance, we logically see it as an alternative to the previous instance, but in fact we keep it in our sub-sample. This is called **exchange sampling.**
- The aggregation or Model training is explained in below example. here we have select each boot instance and try to make classifier here we have used decision tree classifier for each sample. Update the status of result of classifier based on actual and predicted value.



Bagging Code Implementation

Now, Let's move to the implementation section. here we can use sklearn inbuilt iris data set to train a model.

Open in Colab

Import All the required packages from sklearn

In [0]:
```
# Import All the required packages from sklearn
import numpy as np
from sklearn import model_selection
from sklearn.ensemble import BaggingClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.datasets import load_iris
```

Load data

In [0]:
```
iris = load_iris()
X = iris.data
Y = iris.target
```

Split data in training and testing set

Bagging.ipynb hosted with ❤ by GitHub view raw

As you can see, two out of five models are 100% accurate. Models 3 and 9 lack performance as they only have accurate predictions at 90.9999 % and 70.00 %. However, if we average the results, we get an average of accurate predictions of 95.55 %, which is pretty good.

Thanks for reading...!!! Happy Learning...!!!

References :

1. https://stats.stackexchange.com/questions/350897/stacking-without-splitting-data
2. https://www.linkedin.com/pulse/do-you-want-learn-stacking-blending-ensembling-machine-soledad-galli/
3. https://en.wikipedia.org/wiki/Bagging