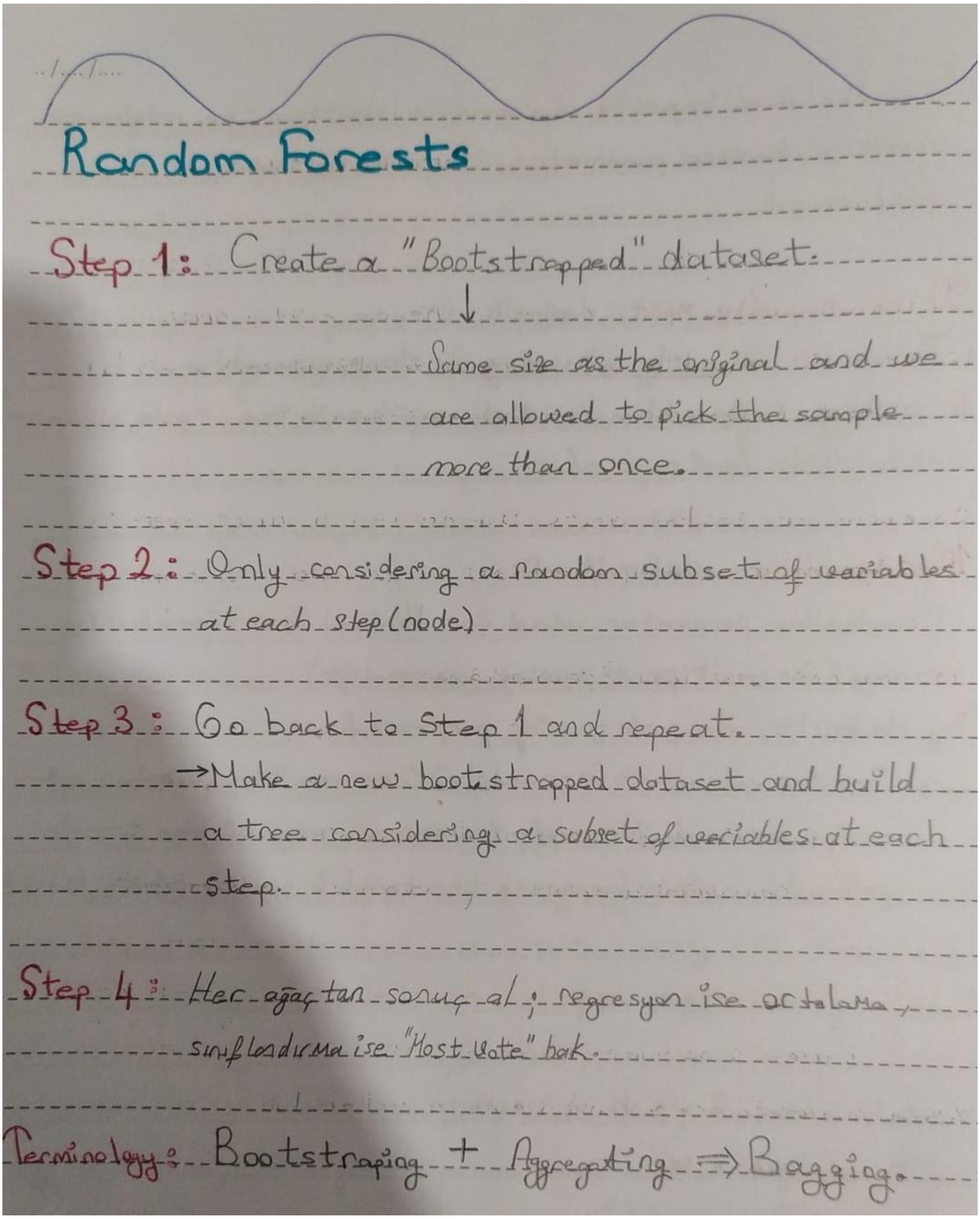


Random Forests

Step 1: Create a "Bootstrapped" dataset.

↓

Same size as the original and we are allowed to pick the sample more than once.

Step 2: Only considering a random subset of variables at each step (node)

Step 3: Go back to Step 1 and repeat.

→ Make a new bootstrapped dataset and build a tree considering a subset of variables at each step.

Step 4: Her ağaçtan sonuç al; regresyon ise ortalama, sınıflandırma ise "Most Vote" bak.

Terminology: Bootstraping + Aggregating ⇒ Bagging.

Feature Importance'lere Karar Verme:

1.1: Default Scikit-learn's (impurity-based) Feature Importance ⇒

Based on impurity, in case of Classification ⇒ Gini impurity

" " " Regression ⇒ Variance

When training a tree we can compute how much each feature contributes to decreasing the weighted impurity.

* Of course for Random Forests we are talking about averaging the decrease in impurity over trees. (Mean Decrease in Impurity)

! Bu yöntemin en büyük dezavantajı:

✗ Biased approach, as it has a tendency to inflate (şişirmek, yüceltmek) the importance of continuous features or high-cardinality categorical variables.

High cardinality anlamı → (too many unique values.)

1.2: Permutation Based Feature Importance (Mean Decrease Accuracy)

Can be used to overcome drawbacks of default feature importance with mean impurity decrease.

Nasıl uygularız?

a) Bir model eğitelim

b) Test kümesinde performansı ölçelim.

c) Değişkenleri tek tek alıp, sadece o değişkeni rastgele shuffle edelim.

d) Test kümesindeki performansa tekrar bakalım

c) Shuffle ve non-shuffle arası farkın (RMSE, Accuracy) en fazla olduğu değişken en önemlidir.

★ Impurity-based (Default Sklearn) importances are computed on training set; Permutation-based is computed on test set or if you want on training set.

1.3 Shap values based Importance

It uses the Shapley values from game theory to estimate the how does each feature contribute to the prediction.

↳ Features with large absolute Shapley values are important.

Important : Shap shows the contribution or the importance
of each feature on the prediction of the model,
it does not evaluate the quality of the prediction
itself.

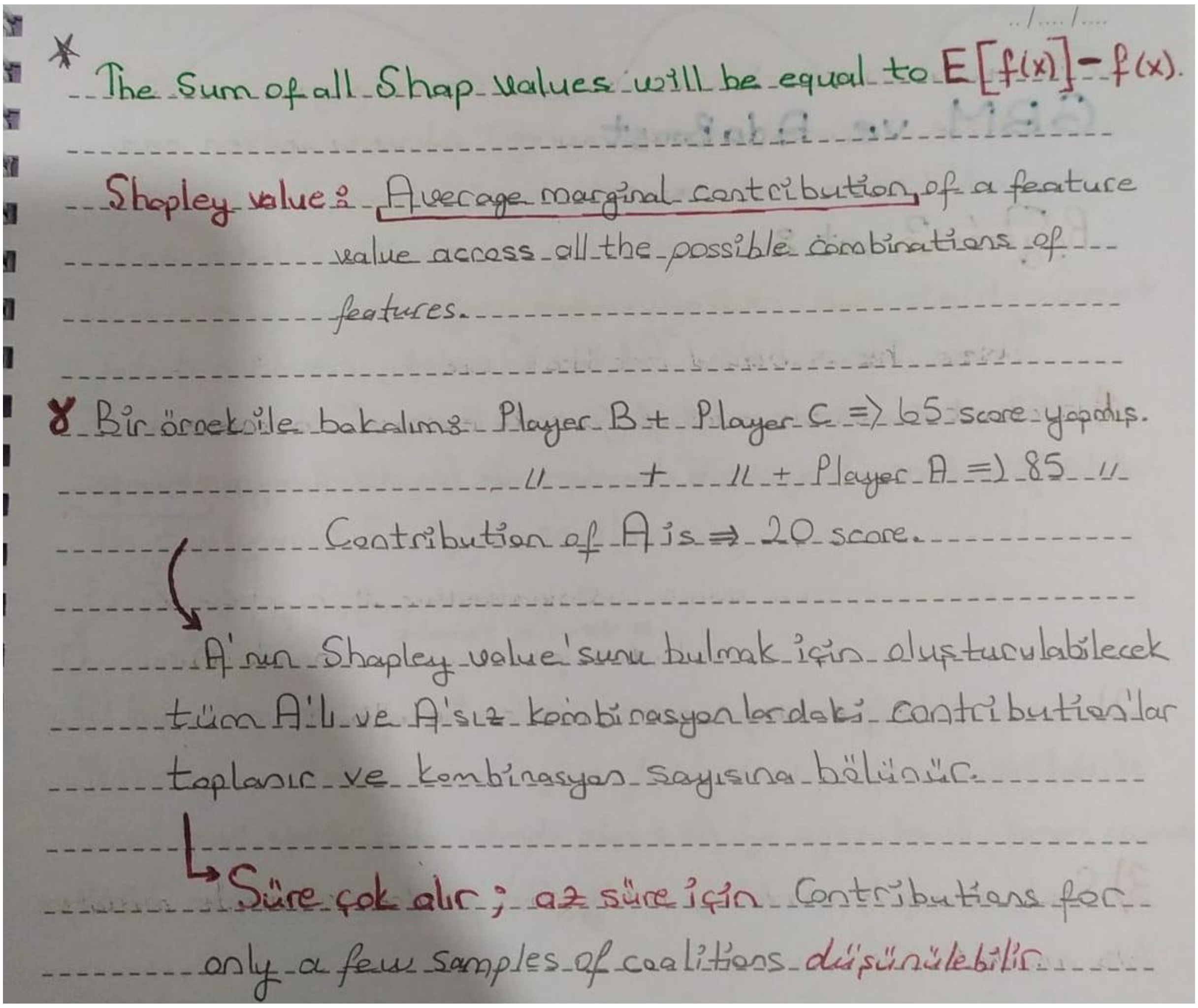
Bu değerler nasıl belirleniyor?

Shap disclose the individual contribution of each feature on the output of the model, for each observation.

Let's say : "x" is the chosen observation
"f(x)" is the predicted value of the model
E[f(x)] is expected value of the target variable
*(↑ The mean of all predictions)

» The absolute Shap value shows us how much a single feature affects PREDICTION.

» In each observation Shap values will differ.



* The Sum of all Shap values will be equal to $E[f(x)] - f(x)$.

Shapley value: Average marginal contribution of a feature value across all the possible combinations of features.

Bir örnek ile bakalım: Player B + Player C => 65 score yapmış.
" + " + Player A => 85 "
Contribution of A is => 20 score.

A'nın Shapley value'sunu bulmak için oluşturulabilecek tüm A'lı ve A'sız kombinasyonlardaki contribution'lar toplanır ve kombinasyon sayısına bölünür.

→ Süre çok alır; az süre için Contributions for only a few samples of coalitions düşünülebilir.