**Decision Tree**

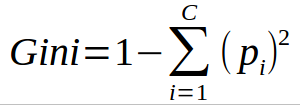
Ein Entscheidungsbaum (Decision Tree) ist ein weit verbreitetes maschinelles Lernverfahren, das sowohl für Klassifikations- als auch für Regressionsaufgaben verwendet wird. Er ist leicht zu verstehen und zu interpretieren, was ihn zu einem beliebten Werkzeug in der Datenanalyse macht. Hier ist eine detaillierte Erklärung dazu:

**Funktionsweise eines Entscheidungsbaums**

1. **Struktur:** Ein Entscheidungsbaum besteht aus Knoten und Zweigen. Die Knoten repräsentieren Entscheidungen oder Tests an einem Attribut, während die Zweige die Ergebnisse dieser Tests und zu den nächsten Knoten führen. Die Blätter des Baums repräsentieren die endgültigen Entscheidungen oder Vorhersagen.
2. **Bau des Baumes:** Der Baum wird durch wiederholtes Teilen des Datensatzes nach bestimmten Kriterien gebaut. Diese Kriterien basieren darauf, wie gut ein Split die Klassen oder Werte in den Daten trennt. Der Prozess des Teilens wird rekursiv fortgesetzt, bis eine vorgegebene Tiefe erreicht ist oder keine signifikante Verbesserung mehr erzielt wird.
3. **Splitting-Kriterien:**
   * **Gini-Index: Ein Maß für die Ungleichheit in einer Verteilung, das verwendet wird, um die "Reinheit" eines Splits zu beurteilen. Ein niedrigerer Gini-Index bedeutet eine homogenere Verteilung.**
   * **Entropie: Ein Maß für die Unsicherheit oder Unordnung, das in der Informationstheorie verwendet wird. In Entscheidungsbäumen wird die Entropie genutzt, um die Information Gain zu berechnen.**
   * **Information Gain: Der Gewinn an Information durch einen bestimmten Split, berechnet als Differenz der Entropie vor und nach dem Split.**

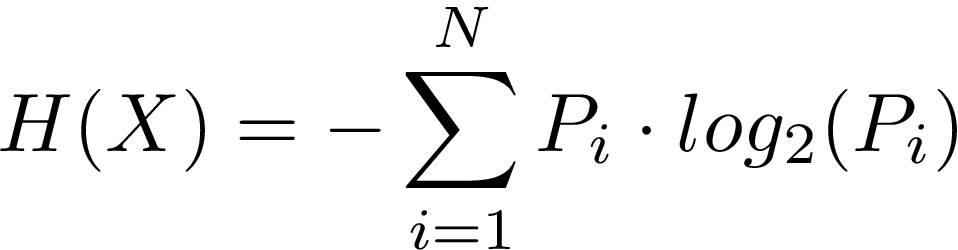
**Mathematische Grundlagen**

**1.Gini-Index**: Der Gini-Index für einen Split in einem Knoten t wird wie folgt berechnet:

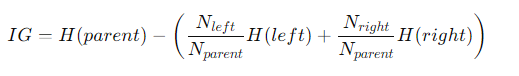


Wobei pi der Anteil der Klasse i am Knoten t ist.

**2.Entropie und Information Gain**: Die Entropie H eines Knoten X wird wie folgt berechnet:



Der Information Gain durch einen Split wird dann berechnet als:



wobei Nleft, Nright und Nparent die Anzahl der Beispiele im linken, rechten und ursprünglichen Knoten sind.

Quellen:

<https://medium.com/@arpita.k20/gini-impurity-and-entropy-for-decision-tree-68eb139274d1>

<https://www.geeksforgeeks.org/decision-tree-introduction-example/>

<https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>

**Random Forest**

Ein Random Forest ist eine erweiterte Methode des Entscheidungsbaums, die durch die Aggregation mehrerer Entscheidungsbäume die Vorhersagegenauigkeit und die Robustheit verbessert. Hier ist eine detaillierte Erklärung der Funktionsweise und der mathematischen Grundlagen, sowie seriöse, frei zugängliche Quellen:

**Funktionsweise eines Random Forest**

1. **Struktur:**
   * **Ein Random Forest besteht aus einer Vielzahl von Entscheidungsbäumen, die unabhängig voneinander auf verschiedenen Bootstraps des Trainingsdatensatzes trainiert werden. Jeder Baum trägt eine Stimme zur Vorhersage bei, und die endgültige Entscheidung wird durch Mehrheitsentscheidung (bei Klassifikation) oder Mittelwertbildung (bei Regression) getroffen.**
2. **Bau des Waldes:**
   * **Bei der Konstruktion eines Random Forests wird für jeden Baum eine zufällige Auswahl von Trainingsdaten (mit Zurücklegen) verwendet, bekannt als Bootstrapping.**
   * **Bei jedem Split in einem Baum wird eine zufällige Untermenge der Merkmale betrachtet, was die Korrelation zwischen den Bäumen reduziert und die Generalisierungsfähigkeit des Modells verbessert.**
3. **Vorteile:**
   * **Höhere Genauigkeit und Robustheit gegenüber Überanpassung (Overfitting) im Vergleich zu einzelnen Entscheidungsbäumen.**
   * **Geringere Anfälligkeit für Ausreißer und Rauschen in den Daten.**

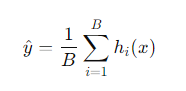
**Mathematische Grundlagen**

1. **Bootstrapping:**
   * **Das Bootstrapping-Verfahren erstellt mehrere Stichproben (mit Zurücklegen) aus dem ursprünglichen Datensatz. Jeder Entscheidungsbaum wird auf einer dieser Stichproben trainiert.**
2. **Random Subspace Method:**
   * **Bei jedem Split wird eine zufällige Untermenge der Merkmale ausgewählt. Dies wird mathematisch durch die Auswahl eines Unterraums der Merkmale beschrieben.**
3. **Mehrheitsentscheidung und Mittelwertbildung:**
   * **Bei der Klassifikation wird die Vorhersage des Random Forest durch die Mehrheitsentscheidung der einzelnen Bäume bestimmt. Mathematisch bedeutet dies:**



Wobei hi(x) die Vorhersage des i-ten Baums für den Eingabewert x ist.

* Bei der Regression wird die Vorhersage durch den Durchschnitt der Vorhersagen der einzelnen Bäume gebildet:



wobei B die Anzahl der Bäume im Wald ist.

Quellen:

<https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/>

<https://towardsdatascience.com/understanding-random-forest-58381e0602d2>