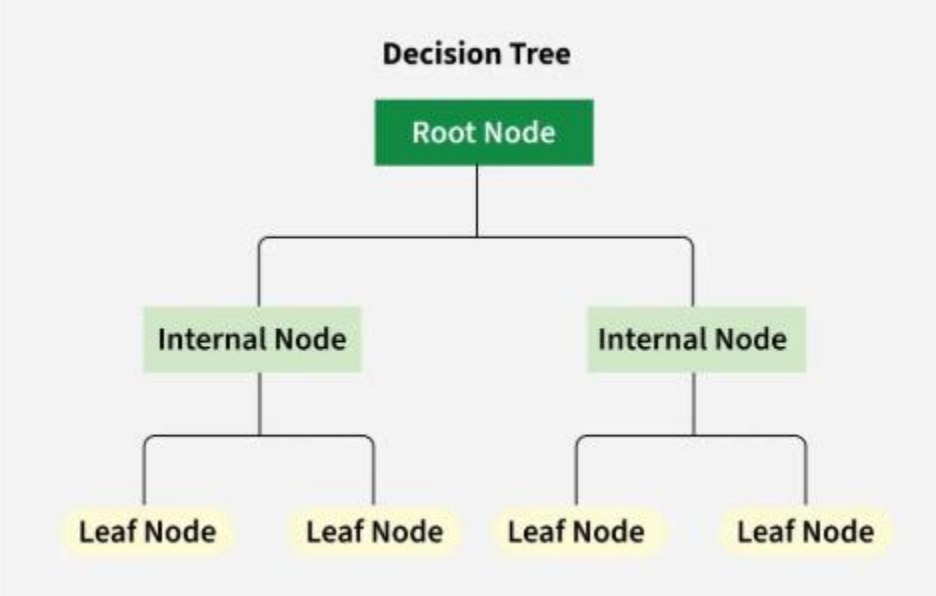
Päätöspuut

Päätöspuu (Decision Tree) on yksinkertainen tietorakenne ja tehokas koneoppimisen työkalu. Päätöspuu muistuttaa tietyllä tapaa vuokaaviota, jossa jokainen sisäinen solmu (internal node) edustaa päätöstä tiettyjen ominaisuuksien (feature) perusteella, jokainen haara kuvaa päätöksen lopputulosta ja jokainen lehtisolmu (leaf node) antaa lopullisen ennusteen. Intuitiivinen rakenne tekee päätöspuista helposti ymmärrettäviä ja ne soveltuvat erilaisiin luokittelu- ja regressio-ongelmiin.



Päätöspuualgoritmeja on useita erilaisia, joista jokaisella on omat erityispiirteensä. Yleisimmät algoritmit ovat: CART (Classification and Regression Tree, ID3 (Iterative Dichotomiser 3, CHAID (Chi-square Automatic Interaction Detector, Random Forest ja Gradient Boosted Trees.

Miten Random Forest toimii?

Random Forest -algoritmi toimii luomalla kokoelman päätöspuita, joita kutsutaan "metsäksi". Kukin puu koulutetaan eri osajoukolla dataa. Algoritmin keskeiset ominaisuudet ovat:

1. **Bootstrap Sampling:** Jokainen puu koulutetaan satunnaisella osajoukolla dataa. Tämä varmistaa, että jokainen puu näkee hieman erilaisen version datasta, mikä lisää monimuotoisuutta.
2. **Feature Randomness:** Puu haarautuminen sisäisessä solmussa tehdään käyttämällä satunnaista määrää ominaisuuksista. Eli jokaisessa sisäisessä solmussa ei käytetä kaikkia ominaisuuksia. Tämä vähentää yksittäisten puiden välistä korrelaatiota ja tekee mallista kestävämmän.
3. **Aggregation:** Luokittelutehtävissä Random Forest yhdistää kaikkien puiden ennusteet enemmistöpäätöksen avulla. Regressiotehtävissä ennusteet yhdistetään laskemalla keskiarvo. Tämä yhdistelmälähestymistapa vähentää overfittingiä ja parantaa tarkkuutta.

Algoritmin edut

Random Forest -algoritmin etuja ovat:

1. Koska Random Forest yhdistelee useita päätöspuita se tasoittaa ”kohinaa” ja vähentää overfitting riskiä, joka on yleistä yksittäisissä päätöspuissa.
2. Algoritmi on vähemmän herkkä datan muutoksille. Pienet vaihtelut datassa eivät yleensä muuta merkittävästi ennusteita.
3. Random Forest on monipuolinen ja toimii sekä luokittelu- että regressiotehtävissä ja käsittelee hyvin sekä numeerisia että kategorisia muuttujia.
4. Random Forest tarjoaa myös tietoa siitä, mitkä ominaisuudet ovat ennustamisen kannalta merkittäviä. Tämä tekee algoritmista hyödyllisen työkalun myös muiden mallien ominaisuuksien valinnassa.

Random Forestia on ilmeisesti laajalti käytössä rahoitusalalla, terveydenhuollossa ja markkinoinnissa. Algoritmi on erityisen hyödyllinen, kun tavoitteena on maksimoida ennustetarkkuus ja minimoida overfitting-riski. Random Forest -algoritmilla on myös joitakin rajoituksia. Algoritmi voi esimerkiksi olla laskennallisesti raskas suurilla aineistoilla tai silloin, kun puita on ”metsässä” paljon.