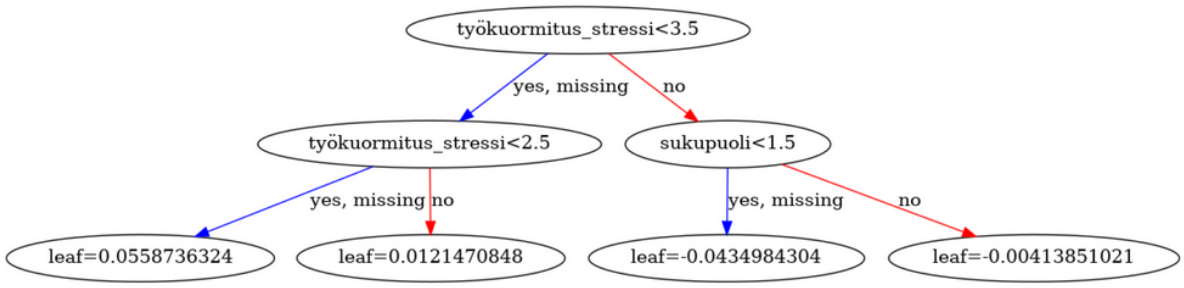
**LIITE 2. XGBoost-menetelmä**

XGBoost (eXtreme Gradient Boosting) on ohjatun koneoppimisen menetelmä, joka soveltuu sekä regressio- että luokitteluongelmien ratkaisemiseen. Menetelmä sovittaa dataan joukon päätöspuita ja muodostaa malliennusteen näiden puiden yhdistelmänä. XGBoost-algoritmi on vastaaviin muihin algoritmeihin verrattuna suorituskykyinen ja sen opettaminen on suhteellisen nopeaa. Nämä ominaisuudet ovat lisänneet sen suosiota ohjatun koneoppimisen sovelluksissa.

Kuvassa on esitetty malli yhdestä mahdollisesta päätöspuusta, jossa on kaksi kaksi alipuuta ja neljä lehtisolmua. Puun jokaisessa sisäsolmussa tehdään vertailu piirremuuttujan arvon ja solmuun liittyvän vertailuarvon välillä ja päätetään sen mukaan, kumpaan haaraan edetään. Kun päädytään lehtisolmuun, saadaan luokiteltavalle näytteelle lehden sisältämä ennustearvo.

**KUVIO. Esimerkki opetetun XGBoost-mallin sisältämästä yksittäisestä päätöspuusta**




Yksittäinen päätöspuu on heikko luokittelija, mutta yhdistämällä useampia heikkoja luokittelijoita samaan malliin voidaan muodostaa vahva luokittelija. Tätä kutsutaan tehostamiseksi (boosting). XGBoost-menetelmässä päätöspuita lisätään inkrementaalisesti siten, että seuraava päätöspuu opetetaan edellisten puiden tekemien virheiden perusteella siirtymällä luokitteluvirhettä kuvaavan kustannusfunktion gradientin suuntaan (gradient boosting). Lopullisen mallin ennuste saadaan laskemalla yksittäisten puiden antamat ennustearvot yhteen.

XGBoost-mallin opettamisessa hyödynnetään tyypillisesti ristivalidointia eli erillisiä opetus- ja validointinäytejoukkoja. Näin voidaan välttää mallin mahdollinen ylisovittuminen eli se, että mallista tulee liian monimutkainen ja se mallintaa myös opetusdatan sisältämää satunnaista kohinaa. Ylisovittunut malli sopii opetusdataan todella hyvin, mutta sen ennustevirhe opetusdatan ulkopuoliselle datalle on suuri.

Opetettu XGBoost-malli mahdollistaa myös siinä käytettyjen piirremuuttujien tärkeyden vertailun. Piirremuuttujan tärkeys muodostetaan yhdelle päätöspuulle siten, että lasketaan sen jokaisen kyseistä piirremuuttujaa käyttävän solmun aiheuttama parannus luokittelun suorituskykyyn painotettuna solmuun liittyvällä opetusnäytteiden määrällä. Koko mallin osalta piirremuuttujan tärkeys saadaan sitten keskiarvoistamalla yli kaikkien mallin puiden antamien tärkeysarvojen. Piirremuuttujien tärkeydet esitetään suhteellisina arvoina, eli niiden summa yli mallin kaikkien piirteiden on yksi.