Random forest

- Yksi maailman suosituimmista koneoppimis algoritmeista
- Laajennus päätöspuusta -> periaate sama
- Metsä koostuu puista
- Metsälle määritellään puiden määrä->mitä enemmän puita sen tarkempi malli
- Data jaetaan pienempiin osiin(bagging), sekoitetaan ja annetaan puille
- Jokainen puu saa oman random datan jolla se treenaa
- Kun koulutus on valmis testataan perinteisesti
- Jokaiselle puulle saadaan tarkkuus lasketaan tarkkuuksien keskiarvo josta saadaan metsän tarkkuus

- Ennustuksessa jokainen puu tekee oman ennustuksensa(äänestää)
- Preddict bagging
- Enemmistö voittaa
- Lisäksi voidaan mainita ennustuksen luotettavuus
- Esim jos kymmenestä puusta 8 antaa vastaukseksi 1 ja kaksi puuta 0
- ->”oikea” vastaus 1 80% luotettavuudella

- Puulle määritellään aloitus node juuri (root) ja syvyys (depth)
- Yritetään valita mahdollisimman hyvin datan jakava root
- Jos saadaan vastaus ennen kuin saavutetaan max depth luodaan leaf ja ollaan tyytyväisiä
- Lopetetaan aina kun saavutetaan max syvyys, vaikkei vastaus olisikaan varma

Edut

- Sopii sekä classifier-ongelmiin että regressio ongelmiin
- Toimii kohtuullisesti “sotkuisella” datalla
- Isompi data ei “overfittaa”(ylikouluta) mallia ihan niin helposti
- Pystyy käsittelemään tehokkaammin isompaa dataa kuin yksittäinen puu
- Antaa luotettavampia ennustuksia kuin yksittäinen puu

Haitat

- Regressio-ongelmissa huonompi kuin luokittelussa
- Saattaa overfitatta mallin
- Black box Käyttäjällä vähän kontrollia mitä malli tekee

Missä käytetään

- talous /pankkisektorilla luokitellaan osakkeita ja asiakkaita
- Konenäkö hahmontunnistus Microsoft Kinect
- Äänentunnistus
- Tekstintunnistus kategorisointi
- Lääketiede: Sairauksien luokittelua oireiden mukaan