# Päiväkirja

## Contents

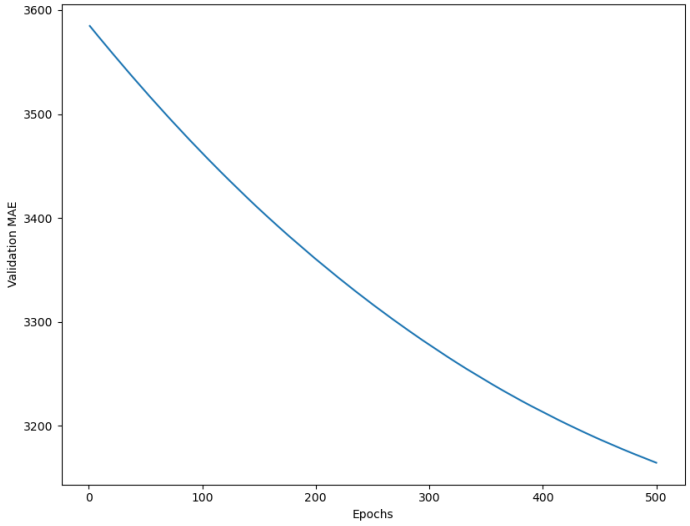
# 1 23.10.2023

Ajoin tosiaan ekat testit. Mukana oli kaikki vektorit, joille oli väestörakennedata ja vuosikate jossain määrin eheänä. Niitä tuli 6800. Ajoin ensimmäiseksi 100 epokkia 4-fold validationilla. Mallissa oli 2 keskikerrosta leveydellä 64 ja aktivaatiofunktio oli relu. Tulos näytti MAE (Mean Value Error) olevan 7 000 000.

Ajoin toisen testin vielä illalla. Kirjan esimerkkiä taas tapaillen tällä kertaa tutkittiin epokkien määrän vaikutusta. Muuten sama setuppi, mutta epokkeja 500 ja kerätään tulokset joka välissä analyysiä varten. Testit ottivat n 1,5h. Unohdin ottaa graafit ylös tällä kertaa, mutta niissä näkyi loiveneva lasku jostain 7400 $\rightarrow$ 6800. Odotin että käyrä olisi lähtenyt nousuun ja olisi löytynyt selkeä ylisovituksen alku, mutta eipä näkynyt. Käyrä kuitenkin loiveni siihen malliin, että ei kovin paljoa parempaa tulosta saisi pelkällä epokkien määrän kasvattamisella.

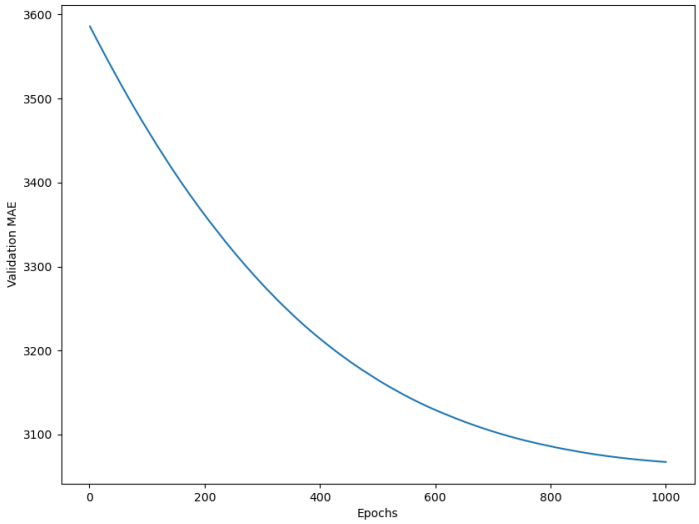
# 2 27.10.2023

Konsultoituani chatGPT:tä ja muistettuani mm, että voisi BKT:n lisätä muuttujaksi. Muovasin vähän hommia. Nyt pyörii vektoreina 3 peräkkäistä vuotta BKT:n kanssa. Otin myös suurimpien vuosikatteiden kunnat pois. Viimeisen vuoden vuosikate tulosteena. Vektorien määrä nyt n 4500. Pitäydyin vielä 4-foldissa, mutta lisäsin yhden kerroksen keskelle lisää ja asetin kerrosten leveydeksi 80. Tulokset toivottavasti tunnin sisään.




Käyrän muoto on hyvin vastaava kuin aiemmalla. Selkeähköä laskua, mutta jänskättää laittaa epokkeja isoja määriä k-fold validationilla, kun sieltä vuotaa tietoa kaiketi (päivitys, ei kaiketi vuodakaan, en tiiä). Vuosikatteiden keskiarvo on tällä skaalalla 3440, joten tässä ei vielä taida olla liikaa hurraamista. Tämä testi söi 45 min. Seuraavaan testiin kerroksien leveydeksi 120 ja epokkeja 1000. Testi pyörimään ja lounaalle.

2 tuntia myöhemmin



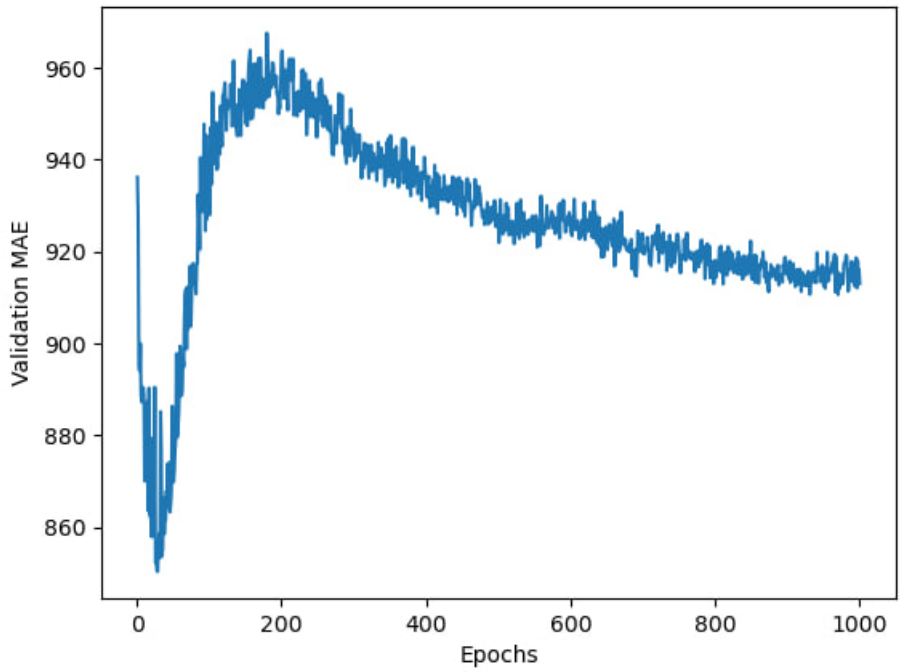

Ei vieläkään ylisovitusta, eikä vieläkään virhe alle 3000. Muita juttuja pitäis keksiä. Onneksi niitä on tyrkyllä kanssa. Latailin vielä noita työssäkäyntitilastoja talteen oikeanmuotoisina exceleinä. Voisin askarrella ne vielä tietokantaan, mutten varmaan enää tänää aja enempää verkon koulutuksia läpi. Tuntuu, että toistaiseksi isoin muuttuja oli se, että otti isoimmat kunnat pois laskuista. Positiivista on, että asiat ovat jo testailtavissa.

# 3 2.1.2024

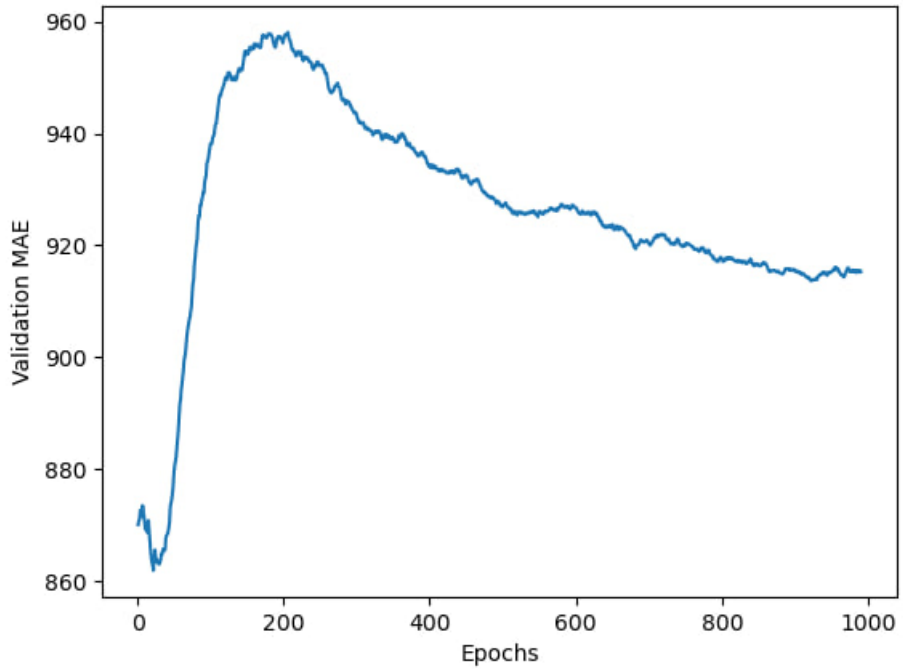
Tietokantaan uudistuksia. Kuntakohtaiset taulut poistettu ja tilalla 6 taulua. Lisäsin vielä väkiluvulle taulun, minkä avulla voi jakaa kuntia näppärästi kokoluokkiin

# 4 9.1.2024

Uusia ajoja uudella datalla, kasa hienosäätöä. Tein rajaukseksi tähän testiin kunnat, joissa keskimäärin max 15k asukasta. 4 foldia, 1000 epookkia. Sisään kaikki saatavilla oleva data kahdelta vuodelta ja ulos ennuste kolmen seuraavan vuoden vuosikatteesta. Kone puksutti reilun vuorokauden ja antoi seuraavaa ulos:




Lisäksi tulosteena myös siistitympi käyrä samasta testistä, ilman 10 ensimmäistä epookkia:




Käyrä näyttää tutummalta kirjan esimerkkiin nähden ja nyt on selvästi rel-

evantimpaa dataa mukana. Ylisovittaminen alkaa nopeasti. Tulokset ovat lupaavampia, mutta vielä ollaan kaukana maalista. Nyt MAE-minimi on jossain 860 paikkeilla. Tämä on siis 3-ulotteisen vuosikatteen virhevektorin mitta. Eli keskimääräinen virhe vuosikatteessa on jossain 520k tienoilla. Tämä oletuksella, että yhdessä kunnassa peräkkäisten vuosien vuosikatteet ovat lähellä toisiaan. Tällöin virhevektorin $\overline{v}$ normista saa

$$|\overline{v}| = \sqrt{v_1^2 + v_2^2 + v_3^2} \approx \sqrt{v_1^2 + v_1^2 + v_1^2}$$

$$\rightarrow v_1 = \frac{|\overline{v}|}{\sqrt{3}}$$

Tämän lisäksi järjestelin koodia monesta paikkaa. Lisäilin pieniä tulosteita konsoliin kertamaan datasta ja kokeen viemästä ajasta. Refaktoroin sieltä täältä ja uudelleenjärjestin kansiorakennetta. Pyrin myös viemään enemmän ja enemmän asioita muuttujiksia mainissä kutsuttavaan funktioon. Nyt siellä on Epookit, foldien määrä ja verkon rakenne aktivaatiofunktioineen. Ajatuksena saada myös kuntien luokittelu muuttujaksi (esim haluaako kunnat, joissa $< N$ asukasta vai jotain väliltä $[x, y]$).

Hienosäätöä lienee mahdollista tehdä loputtomiin, mutta tällä hetkellä se tuntuu mukavammalta kuin ajatus pilvipalveluihin tutustumisesta tai uuden datan etsimisestä.