OULUN AMMATTIKORKEAKOULU

OPPIMISPÄIVÄKIRJA GOOGLE ADVANCED DATA ANALYTICS - SERTIFIKAATIN SUORITTAMISESTA

Joonas Valle
Opinnäytetyö
Syksy 2025
Tietojenkäsittelyn tutkinto-ohjelma
Oulun ammattikorkeakoulu

TIIVISTELMÄ

Oulun ammattikorkeakoulu
Tietojenkäsittelyn tutkinto-ohjelma

Tekijä: Joonas Valle
Opinnäytetyön otsikko: Oppimispäiväkirja Google Advanced Data Analytics -sertifikaatin suorittamisesta
Työn valmistumislukukausi ja -vuosi: Kevät 2026
Sivumäärä: 34

Opinnäytetyössä suoritettiin Google Advanced Data Analytics -sertifikaatti, josta laadittiin päiväkirjamuotoinen raportti. Päiväkirja kuvaa sertifikaatin suorittamisen eri vaiheet sekä esittelee kurssien sisällöt kurssikohtaisesti. Opinnäytetyön tavoitteena oli kuvata sertifikaatin suorittamista ja sen hyötyjä opiskelijalle sekä edistää opiskelijan osaamista data-analytiikassa.

Opinnäytetyön tietoperusta on rajattu sertifikaatin keskeisiin sisältöihin. Pääteemana on Python data-analytiikassa. Tietoperusta käsittelee myös data-analytiikkaa, Pythonin keskeisiä kirjastoja, tilastollista analyysiä sekä koneoppimista.

ABSTRACT

Oulu University of Applied Sciences
Bachelor of Business Information Systems

Author: Joonas Valle
Title of thesis: Learning Diary on the Completion of the Google Advanced Data Analytics Certificate
Term and year when the thesis was submitted: Spring 2026
Number of pages: 34

This thesis involved completing the Google Advanced Data Analytics Certificate, which was documented in the form of a learning diary report. The diary describes the different stages of completing the certificate and presents the course contents on a course-by-course basis. The objective of the thesis was to describe the completion of the certificate and its benefits for the student, as well as to enhance the student's competence in data analytics.

The theoretical framework of the thesis was limited to the key contents of the certificate. The main theme is Python in data analytics. In addition, the theoretical framework covers data analytics, key Python libraries, statistical analysis, and an introduction to machine learning.

SISÄLLYS

1 JOHDANTO

Tämän päiväkirjamuotoisen opinnäytetyön tarkoitus on kuvata Courseran sertifikaatin suorittamista, sen sisältöä sekä hyötyä raportin kirjoittajalle. Opinnäytetyön ohessa suoritetaan Google Advanced Data Analytics -sertifikaatti, joka löytyy Coursera-oppimisalustalta. Sertifikaatin suorittamisesta laaditaan päiväkirja, joka kertoo sertifikaatin eri osista ja niiden sisällöstä. Sertifikaatti koostuu seitsemästä eri kurssista. Kurssien päätyttyä niistä kirjoitetaan päiväkirjaan myös yhteenveto.

Kurssit suoritetaan oppimisalustalla itsenäisesti omaan tahtiin. Kurssien sisältö rakentuu yleensä 4–6 moduulista, jotka sisältävät opetusta eri muodoissa. Moduulit koostuvat opetusvideoista, lukumateriaalista, tehtävistä, laboratorioharjoituksista, testeistä ja moduulien loppukokeista. Opetusvideoissa esiintyy yleensä alan ammattilaisia, jotka kertovat aiheesta sekä omista kokemuksistaan alalla. Sertifikaatti on tasoltaan edistynyt, ja on jatkoa aloittelijatason Google Data Analytics -sertifikaattiin, jonka suoritin aiemmin syksyllä osana tutkinto-ohjelmaa. Sertifikaatin tarjoaa ja toteuttaa Google Career Certificates.

Raportin pääteemana on Pythonin hyödyntäminen data-analytiikassa. Raportissa käsitellään myös data-analytiikan työprosessia, keskeisiä Python-kirjastoja sekä Pythonin hyödyntämistä tilastollisessa analyysissä ja koneoppimisessa.

Valitsin tämän sertifikaatin opinnäytetyöhöni, koska sen aihe kiinnosti minua jo aiemmin suoritetun sertifikaatin myötä. Aihe tuntui myös helposti lähestyttävältä, koska olin perehtynyt siihen jo aiemmin perusteellisesti. Sertifikaatin aikana käytettävän Python-ohjelmointikielen opettelu vaikutti myös kiinnostavalta sekä tulevaisuuden kannalta hyödylliseltä.

2 SANASTO

RACI-taulukko – Vastuunjakomatriisi roolien ja vastuiden selkeyttämiseen projektityöskentelyssä (Responsible, Accountable, Consulted, Informed).

PACE-malli – Nelivaiheinen toimintamalli, joka ohjaa työskentelyä suunnittelusta toteutukseen (Plan, Analyze, Construct, Execute).

Jupyter Notebook – Selainpohjainen kehitysympäristö, jossa voidaan kirjoittaa ja suorittaa ohjelmakoodia vaiheittain.

EDA (Exploratory Data Analysis) – Tutkiva data-analyysi on aineiston systemaattista tarkastelua ja visualisointia.

Z-arvo – Standardoitu tunnusluku, jota laskemalla havaitaan poikkeamat.

Khiin neliö -testi – Tilastollinen menetelmä, jolla voi mitata kuinka paljon havaittu data poikkeaa siitä, mitä olisi odotettu.

ANOVA – Menetelmä, jolla vertaillaan useamman ryhmän keskiarvoja ja tarkistetaan, ovatko niiden erot todennäköisesti oikeita vai sattumaa.

Naive Bayes -malli – Todennäköisyyspohjainen luokittelumenetelmä, joka perustuu Bayesin teoreemaan.

K-means-klusterointi – Ohjaamaton koneoppimismenetelmä, jossa data jaetaan K ryhmään niiden keskipisteiden perusteella.

XGBoost-malli (Extreme Gradient Boosting) – Puupohjainen koneoppimismenetelmä, joka perustuu gradient boosting -tekniikkaan.

3 PYTHON DATA-ANALYTIIKASSA

3.1 Data-analytiikka ja sen työprosessi

Data-analytiikalla tarkoitetaan teorioiden, teknologioiden, välineiden ja prosessien kokonaisuutta, jonka avulla datasta pyritään tuottamaan ymmärrystä ja merkityksellisiä havaintoja. Se ei rajoitu pelkkään datan tarkasteluun, vaan sisältää systemaattisen lähestymistavan datan käsittelyyn, analysointiin ja tulkintaan. Data-analytiikan tavoitteena on tuottaa tietoa, jota voidaan hyödyntää päätöksenteossa ja toiminnan kehittämisessä. (Sarker 12.7.2019.)

Data-analytiikka etenee vaiheittaisena prosessina, jolla pyritään ratkaisemaan ennalta määritelty ongelma. Prosessi alkaa aineiston keruulla, jonka jälkeen data puhdistetaan ja esikäsitellään analyysiä varten. Analyysivaiheessa hyödynnetään erilaisia menetelmiä havaintojen ja yhteyksien tunnistamiseksi, minkä jälkeen tulokset tulkitaan ja raportoidaan päätöksenteon tueksi. (Thavali 16.10.2025.) Organisaatio voi esimerkiksi tarkastella asiakastyytyväisyyteen vaikuttavia tekijöitä analytiikan avulla.

Data-analytiikan merkitys on kasvanut digitalisaation myötä, ja datan hyödyntämisestä on tullut keskeinen osa organisaatioiden päätöksentekoa ja kilpailukykyä (World Economic Forum, 2025). Data-analytiikan soveltaminen käytännössä perustuu ohjelmointikieliin ja analytiikkatyökaluihin, joiden avulla dataa voidaan käsitellä ja mallintaa tehokkaasti. Näistä erityisesti Python on vakiinnuttanut asemansa yhtenä keskeisimmistä data-analytiikan työkaluista (Clement 8.10.2024).

3.2 Python data-analytiikan työkaluna

Miksi Python on niin laajasti suosittu ohjelmointikieli sekä yleisesti että erityisesti data-analytiikassa? Pythonin suosiota voidaan tarkastella kolmen keskeisen tekijän kautta: selkeä ja helposti omaksuttava syntaksi, monipuoliset käyttömahdollisuudet sekä laaja ja aktiivisesti kehittyvä kirjastoekosysteemi

(Saari & Ahvonen 20.1.2026). Data-analytiikassa tämä tarkoittaa, että analyytikko voi keskittyä itse dataan eikä kielen teknisiin yksityiskohtiin. Kielen selkeä rakenne ja vähäinen muodollisuus tekevät koodista helposti luettavaa myös niille, joilla ei ole pitkää ohjelmointitaustaa. Pythonin luettavuus ei hyödytä ainoastaan yksittäistä analyytikkoa, vaan se tukee tiimityötä ja helpottaa viestintää eri sidosryhmien kanssa. Pythonin koodi muistuttaa monin osin luonnollista kieltä. (Verbina 24.9.2025.) Pythonin syntaksin yksinkertaisuutta ja selkeää rakennetta havainnollistetaan kuvassa 1.

```
# Tallennetaan myyntiluvut
myynti = [120, 150, 130, 170]

# Lasketaan keskiarvo
keskiarvo = sum(myynti) / len(myynti)

print(keskiarvo)
```

```
142.5
```

Kuva 1. Yksinkertainen esimerkki Python-koodista, jossa lasketaan myyntilukujen keskiarvo.

Pythonin suosiota selittää myös sen avoimen lähdekoodin luonne ja laaja, aktiivinen kehittäjäyhteisö. Pythonia on kehitetty 1990-luvun alusta lähtien, jolloin alankomaalainen Guido van Rossum julkaisi kielen ensimmäisen version (Hayriye 3.11.2024). Avoin lähdekoodi mahdollistaa sen, että kehittäjäyhteisö voi tarkastella, testata ja kehittää ohjelmointikieltä edelleen, mikä lisää sen luotettavuutta ja vakautta. Vuosikymmenten kehitys on luonut pohjan laajalle kirjastotarjonnalle, mikä näkyy erityisesti data-analytiikan työkalujen monipuolisuudessa.

Pythonin käyttäjäyhteisö muodostaa laajan kansainvälisen tukiverkoston, joka tarjoaa keskustelufoorumeita, oppimateriaaleja ja avoimen lähdekoodin projekteja. Yhteisön aktiivisuus näkyy muun muassa nopeana tiedonvaihtona ongelmatilanteissa, mikä ilmenee Python-aiheisten kysymysten ja vastausten runsaana määränä Stack Overflow -sivustolla. (Verbina 24.9.2025.)

Kansainvälinen yhteisö edistää tiedon ja käytäntöjen leviämistä niin aloittelijoiden kuin kokeneiden ammattilaisten keskuudessa. Python-yhteisö järjestää myös monia tapahtumia vuoden ympäri eri puolilla maailmaa, joista tunnetuimpia ovat PyCon-konferenssit (Python 2026).

3.3 Keskeiset kirjastot

Pythonin analytiikkakäytön keskeinen vahvuus on sen laaja kirjastoekosysteemi, joka kattaa koko data-analytiikan prosessin datan käsittelystä mallintamiseen ja visualisointiin. Valmiit, laajasti käytetyt kirjastot tehostavat työskentelyä ja vähentävät tarvetta käyttää useita erillisiä ohjelmistoja. (Karl 31.5.2024.) Vaikka Python ei perustu maksimaaliseen suorituskykyyn, sen tarjoamat työkalut ja kirjastot nopeuttavat analyysien toteuttamista (Saari & Ahvonen 20.1.2026). Python-kirjasto on kokoelma valmiita toimintoja, joita voidaan hyödyntää tiettyyn käyttötarkoitukseen ohjelmointityössä. Seuraavaksi tarkastellaan keskeisiä Python-kirjastoja, joita hyödynnetään laajasti data-analytiikassa: NumPy, pandas, Matplotlib, seaborn sekä scikit-learn.

NumPy – numeerisen laskennan perusta

NumPy on Python-kirjasto, joka tarjoaa tehokkaita työkaluja numeeriseen laskentaan ja taulukkomuotoisen datan käsittelyyn. Sen ydinrakenne on moniulotteinen taulukko (array), jonka avulla suuria aineistoja voidaan käsitellä tehokkaasti. (NumPy s.a.)

NumPyn tehokkuus perustuu sen tapaan tallentaa data yhtenäiseen muistirakenteeseen sekä sen taustalla toimiviin C-kielellä toteutettuihin algoritmeihin. Tämän ansiosta laskutoimitukset voidaan suorittaa ilman Pythonin tulkintakerroksen aiheuttamaa lisäkuormaa, mikä parantaa suorituskykyä erityisesti suurilla aineistoilla. (McKinney 2022, 4.) Sen integraatio pandas-kirjaston kanssa sujuvoittaa datan käsittelyä ja analysointia. Tämän vuoksi NumPyä pidetään yhtenä keskeisimmistä työkaluista data-analytiikassa. (Almaci 17.12.2023.)

Pandas – datan käsittely ja esikäsittely

Pandas on avoimen lähdekoodin Python-kirjasto, jota käytetään yleisesti aineiston käsittelyyn. Se tukee erityisesti analyysiprosessin alkuvaiheita, joissa aineistoa valmistellaan jatkokäsittelyä ja mallintamista varten. (Coursera Staff 13.11.2025.)

Data-analytiikassa pandas-kirjastoa hyödynnetään erityisesti aineiston esikäsittelyssä, jossa dataa puhdistetaan, suodatetaan ja muokataan analyysin tarpeiden mukaiseksi. Kirjaston avulla voidaan käsitellä puuttuvia arvoja, yhdistää eri lähteistä peräisin olevia aineistoja sekä tuottaa yhteenvetoja ja ryhmittelyjä, jotka tukevat aineiston rakenteen ymmärtämistä. (McKinney 2022, 5.) Näin pandas toimii keskeisenä työkaluna ennen varsinaista tilastollista analyysiä tai koneoppimismallien rakentamista.

Matplotlib ja seaborn – visualisointi

Matplotlib ja seaborn ovat Python-kirjastoja, joita käytetään aineiston visualisointiin. Visualisointi on keskeinen osa analyysiprosessia, sillä graafinen esitystapa tukee datan rakenteen hahmottamista, poikkeamien tunnistamista sekä tulosten viestintää sidosryhmille (Unwin 31.1.2020).

Matplotlib-kirjaston avulla aineistosta voidaan tuottaa kaksiulotteisia kaavioita ja muita visuaalisia esityksiä. Se rakentuu NumPy-kirjaston päälle ja mahdollistaa numeerisen datan esittämisen graafisessa muodossa. Data-analytiikassa Matplotlibia hyödynnetään erityisesti muuttujien välisten suhteiden, jakaumien ja trendien tarkasteluun. Kirjaston avulla voidaan tuottaa esimerkiksi viiva-, hajonta- ja pylväskaavioita sekä histogrammeja, joita voidaan muokata akselimerkintöjen ja muiden visuaalisten ominaisuuksien osalta analyysin tarpeiden mukaisesti. (Day 15.7.2025.)

Seaborn täydentää Matplotlibia tarjoamalla visualisointityökaluja, jotka tukevat tilastollista analyysiä. Kirjasto mahdollistaa esimerkiksi jakaumien, korrelaatioiden ja ryhmittelyjen havainnollistamisen selkeässä muodossa, mikä tukee datan tulkintaa analyysiprosessin aikana. Lisäksi sen yhteensopivuus pandasin kanssa helpottaa aineiston visualisointia. (Seaborn s.a.)

Scikit-learn – koneoppiminen ja mallinnus

Scikit-learn on avoimen lähdekoodin Python-kirjasto, joka erikoistuu koneoppimiseen. Kirjasto tarjoaa valmiita työkaluja muun muassa luokitteluun, regressioon ja klusterointiin. (Scikit-learn s.a.) Se rakentuu NumPyn ja SciPyn kaltaisten tieteellisen laskennan kirjastojen päälle ja toimii saumattomasti yhdessä Pandasin kanssa. Kirjasto tarjoaa myös menetelmiä piirrevalintaan ja ulottuvuuksien pienentämiseen sekä laajan valikoiman mittareita mallien suorituskyvyn arviointiin. (Clark s.a.) Tämä tekee siitä keskeisen työkalun data-analytiikan mallinnusvaiheessa.

3.4 Python tilastollisessa analyysissä ja koneoppimisessa

Pythonin analytiikkakäyttö ulottuu datan käsittelyä pidemmälle tilastolliseen mallintamiseen ja koneoppimiseen. Menetelmien tavoitteena on selittää ilmiöitä, testata oletuksia ja tuottaa ennusteita aineistosta. Seuraavassa kuvataan tiiviisti regressioanalyysi, hypoteesitestaus, mallien arviointi sekä koneoppiminen.

Regressioanalyysi on tilastollinen menetelmä, jolla mallinnetaan selittävien (riippumattomien) muuttujien ja selitettävän (riippuvan) muuttujan välistä tilastollista yhteyttä sekä tuotetaan ennusteita. Menetelmää käytetään sekä ilmiöiden selittämiseen että tulevien arvojen ennustamiseen. On kuitenkin tärkeää huomata, että regressioanalyysi kuvaa muuttujien välistä tilastollista yhteyttä, mutta ei sellaisenaan osoita kausaalisuutta. Yksinkertaisessa regressiossa tarkastellaan yhden selittävän muuttujan vaikutusta, kun taas monimuuttajaregressiossa voidaan huomioida useita tekijöitä samanaikaisesti. Data-analytiikassa menetelmää hyödynnetään esimerkiksi arvioitaessa, miten taustatekijät liittyvät mitattuun lopputulokseen. (Cote 14.12.2021.)

Hypoteesitestaus on tilastollinen menetelmä, jolla arvioidaan, viittaako havaittu ero tai yhteys siihen, että ilmiö on todennäköinen myös perusjoukossa. Testaus perustuu nollahypoteesiin ja vaihtoehtoiseen hypoteesiin, ja päätelmää tuetaan tyypillisesti p-arvon avulla suhteessa ennalta asetettuun merkitsevyystasoon. Analytiikassa hypoteesitestausta voidaan käyttää esimerkiksi kahden ryhmän vertailuun tai sen arviointiin, poikkeaako havaittu tulos satunnaisvaihtelusta.

Tulosten tulkinnassa on kuitenkin tärkeää erottaa tilastollinen merkitsevyys käytännön merkittävyydestä. (Walker 2019, 227–231.)

Mallien arviointi on keskeinen osa sekä tilastollista analyysiä että koneoppimista, sillä mallin toimivuutta tulee tarkastella suhteessa uuteen dataan. Tämän vuoksi aineisto jaetaan usein opetus- ja testiosiin tai hyödynnetään ristivalidointia. Arviointimittarit riippuvat tehtävästä: regressiossa tarkastellaan esimerkiksi ennustevirheitä, kun taas luokittelussa korostuvat tarkkuus ja virhetyyppien hallinta. (Coursera Staff 4.6.2025.)

Koneoppiminen voidaan jäsentää peruskategorioihin sen mukaan, millaista ohjausta mallin oppimisessa on. Ohjatussa oppimisessa mallille annetaan selitettävä muuttuja, ja tavoitteena on ennustaa esimerkiksi jatkuvaa arvoa tai luokkaa. Ohjaamattomassa oppimisessa muuttujaa ei ole, vaan pyritään löytämään rakenteita datasta esimerkiksi ryhmittelyn ja ulottuvuuksien pienentämisen avulla. (Carreiro 15.7.2023.) Pythonin analytiikkaekosysteemi tukee näitä tehtävätyyppejä yhtenäisin työvälinein, mikä helpottaa menetelmien kokeilua, vertailua ja käyttöönottoa osana analyysiprosessia.

Yhdessä tilastolliset menetelmät ja koneoppimisen perusmallit muodostavat työkalupaketin, jonka avulla Pythonilla voidaan sekä selittää aineistoa että tuottaa ennusteita ja päätöksenteon kannalta hyödyllisiä tuloksia.

4 PÄIVÄKIRJA

Suoritan sertifikaatin kurssi kerrallaan järjestyksessä. Laadin muistiinpanoja kurssiin kuuluvien moduulien sisällöstä ja teen niiden pohjalta päiväkirjamerkinnät. Kirjaan myös ylös moduulien testitulokset ja merkitsen ne päiväkirjaan. Teen kurssista yhteenvedon, kun olen suorittanut sen kokonaan.

Sertifikaatti koostuu seitsemästä kurssista, jotka sisältävät noin viisi moduulia per kurssi. Kurssit suoritetaan järjestyksessä seuraavasti:

- Kurssi 1: Foundations of Data Science
- Kurssi 2: Get Started with Python
- Kurssi 3: Go Beyond the Numbers: Translate Data into Insights
- Kurssi 4: The Power of Statistics
- Kurssi 5: Regression Analysis: Simplify Complex Data Relationships
- Kurssi 6: The Nuts and Bolts of Machine Learning
- Kurssi 7: Google Advanced Data Analytics Capstone

4.1 Kurssi 1: Foundations of Data Science

Moduuli 1.

Aloitan sertifikaatin suorittamisen luonnollisesti ensimmäisestä kurssista. Ensimmäisen kurssin ensimmäisessä moduulissa esitellään sertifikaatin sisältö lyhyesti sekä ensimmäisen kurssin, että moduulin sisältö. Sisältö on pääosin lyhyitä videotallenteita sekä lukemista, joiden aiheena on pääasiassa datatieteen konseptit ja data-analyytikon työkuvaan tutustuminen.

Moduuliin sisältyy myös 25 kysymyksen monivalintatesti ja lyhyt data-analytiikkaan liittyvä harjoitus. Monivalintatesti ja harjoitus olivat hyvää kertausta data-analytiikan konsepteihin. Monivalintatestin tarkoitus on määrittää, kuinka valmis olen tähän sertifikaattiin, ja sain siitä tulokseksi 20/25 oikein. Moduuli päättyy 10 kysymyksen pisteytettyyn monivalintakokeeseen, jonka läpäisyyn

vaaditaan vähintään 80 % kysymyksistä oikein. Saan kokeesta ensimmäisellä kerralla täydet 10/10 pistettä.

Moduuli 2.

Toinen moduuli koostuu videotallenteista, lukemisesta, kahdesta monivalintatestistä, harjoituksesta ja pisteytetystä monivalintakokeesta. Moduulin materiaalissa tutustutaan data-analytiikka-alan eri työnimikkeisiin sekä niiden työtehtäviin. Toisena pääkohtana esiintyy data-analytiikan hyödyntäminen nykyteollisuuden eri aloilla. Moduulin harjoituksessa organisoitiin oma datatiimi hyödyntämällä RACI-taulukkoa. Läpäisen taas pisteytetyn monivalintakokeen ensimmäisellä yrittämällä pistein 85 %.

Moduuli 3.

Tämä moduuli on selvästi lyhyempi kuin muut kurssiin kuuluvat moduulit. Se koostuu muutamasta videotallenteesta, lukemisesta, harjoituksesta ja monivalintakokeesta. Aiheena ovat data-analytiikassa ja erityisesti sertifikaatissa käytettävät työkalut. Tekoälyn hyödyntämisestä data-analytiikassa kerrotaan myös lyhyesti.

Harjoituksessa valittiin tekoälytyökalu, jonka avulla harjoiteltiin promptin kirjoittamista. Valitsin työkaluksi ChatGPT:n, koska se oli minulle entuudestaan tuttu. Monivalintakoe menee jälleen ensimmäisellä yrittämällä läpi pistein 87,5 %.

Moduuli 4.

Neljäs moduuli alkaa tutustumisella PACE-malliin ja sen hyödyntämiseen data-analytiikan projektityöskentelyssä. PACE-malli koostuu neljästä vaiheesta (Plan, Analyze, Construct ja Execute), jotka ohjaavat projektin etenemistä suunnittelusta analyysiin ja ratkaisun rakentamisen kautta toteutukseen. Moduuli on huomattavasti pidempi kuin aikaisemmat, ja siihen sisältyy useita lyhyitä monivalintatestejä, videotallenteita, lukemista sekä tehtävä.

Tehtävässä laaditaan projektiehdotus fiktiiviseen työtilanteeseen perustuen ja käytetään valmista mallia apuna. Moduulin pituus ja erityisesti liiallinen näytöltä luettavan tekstin määrä ovat hieman puuduttavia. Tällä kertaa monivalintakoe

vaatii kaksi yritystä: ensimmäisellä kerralla tulos on 67,5 % oikein ja toisella koe menee läpi pistein 97,5 %.

Moduuli 5.

Tämän moduulin aiheena on kurssin päätösprojekti, joka on sertifikaatin ensimmäinen labraharjoitus. Projektin saa valita kolmesta eri aiheesta (Automatidata, TikTok ja Waze), joista valitsen aiheeksi Automatidatan.

Harjoituksessa käsitellään kuvitteellista skenaariota, jossa työskentelen datakonsultointiyritys Automatidatassa, ja yrityksen tehtävänä on kehittää sovellus taksiyritykselle. Minun tehtäväni on käyttää Pythonia datan lataamiseen, tarkasteluun ja järjestämiseen.

Labraharjoitus ei ole liian vaativa, ja se tarjoaa sopivan johdatuksen Python-ohjelmointikielen hyödyntämiseen data-analytiikassa.

Kurssin yhteenveto.

Ensimmäinen kurssi oli luontevaa jatkoa aiemmalle data-analytiikan sertifikaatille, jonka suoritin ennen opinnäytetyöprosessin aloitusta. Aihetta käsiteltiin heti syvemmin, erityisesti Pythonin pariin tutustumalla labraharjoituksen myötä.

Tämä oli minulle myös ensimmäinen labraharjoitus Courserassa, ja se oli mielestäni hyvin opettava ja johdonmukainen. Labraharjoitukset tehdään Jupyter Notebook -ympäristössä. Vaikka sertifikaatti on luokiteltu edistyneeksi, se ei mielestäni ollut ainakaan ensimmäisen kurssin perusteella liian vaativa.

Kurssin videotallenteet olivat Courserassa Googlen tarjoamille kursseille tyypillisesti laadukkaita ja ytimekkäitä. Tekstiosuuksia oli joissakin moduuleissa mielestäni liikaa, enkä opi niistä yhtä hyvin kuin videotallenteista tai harjoituksista. Moduulien pisteytetyt monivalintakokeet ja lyhyemmät, muutaman kysymyksen testit olivat hyvää kertausta ja oman oppimisen testausta. Ensimmäinen kurssi antoi perusteellisen mutta ytimekkään johdannon datatieteen eri konsepteihin.

4.2 Kurssi 2: Get Started with Python

Moduuli 1.

Sertifikaatin suorittaminen jatkuu toisen kurssin ensimmäisellä moduulilla. Moduuli alkaa toisen kurssin sisältöön tutustumisella lyhyesti. Videotallenteet ovat edelleen sujuvia, ja niissä esiintyy uusi henkilö. Sisältöön kuuluu myös lukemista, pari labraharjoitusta sekä monivalintakoe. Ensimmäinen labraharjoitus on lähinnä seuraamista, mutta toisessa harjoituksessa kirjoitetaan jo omaa koodia. Tehtävät sujuvat ongelmitta, koska Pythonin perusfunktiot ovat minulle jo ennestään tuttuja. Läpäisen ensimmäisen monivalintakokeen ensimmäisellä yrityksellä pistein 85 %.

Moduuli 2.

Toinen moduuli koostuu videotallenteista, lukemisesta, muutamasta harjoituksesta sekä monivalintakokeesta. Pääpaino on erityisesti moduulin kolmessa labraharjoituksessa, joissa perehdytään funktioihin, operaattoreihin ja ehtolauseisiin. Ensimmäinen harjoitus on lähinnä seuraamista, mutta kahdessa seuraavassa tuotetaan jo omaa koodia. Harjoituksissa huomaa, etteivät pulmat olekaan niin yksinkertaisia ja niissä pääsee pohtimaan ratkaisuja data-alan ammattilaisen näkökulmasta. Jokaisen harjoituksen lopussa kirjoitetaan myös lyhyt yhteenveto sen sisällöstä, mikä on sopivaa kertausta sekä oman oppimisen jäsentämistä. Monivalintakoe vaatii tällä kertaa kaksi yritystä: ensimmäisellä 70 % oikein ja toisella läpäisen kokeen saamalla 95 % oikein.

Moduuli 3.

Moduuli alkaa parilla videotallenteella, minkä jälkeen se on pitkälti pelkkää harjoitusten työstämistä. Labraharjoitusten aiheena ovat while-silmukat, for-silmukat sekä merkkijonot. Tehtävät ovat aiempaa haastavampia, mutta ne onnistuvat ajan kanssa. Moduulin toisessa harjoituksessa onnistun jumittamaan Jupyter Notebook -ympäristön kokonaan, mutta käynnistämällä kernelin uudelleen tehtävien ratkominen voi jatkua. Neljä harjoitusta putkeen tuntuu hieman puuduttavalta. Lopulta läpäisen monivalintakokeen ensimmäisellä yrityksellä saamalla kysymyksistä 80 % oikein.

Moduuli 4.

Kurssin neljäs moduuli koostuu videotallenteista, lukemisesta ja useista labraharjoituksista. Harjoituksien aiheina ovat listat, tuplet, sanakirjat sekä NumPy- ja pandas-kirjastot. Tehtävien taso pysyy samana, mutta harjoitukset tuntuvat entistä pidemmiltä. Videotallenteita on tällä kertaa todella paljon, mutta ne toimivat sopivina hengästystaukoina loputtomalta tuntuvaan labratyöskentelyyn. Monivalintakoe menee tällä kertaa ensimmäisellä kerralla läpi pistein 90,9 %.

Moduuli 5.

Kurssin viimeisessä moduulissa jatketaan viime kurssilla aloitettua projektia, joka esitetään lyhyesti videotallenteella. Valitsen jälleen aiheeksi Automatidatan, koska valitsin sen viime kurssissakin. Tällä kertaa käytetään jo opittuja funktioita tiedon organisointiin ja rakennetaan DataFrame projektia varten.

Kurssin yhteenveto.

Kurssin sisällön nopeasti skannattuani huomasin heti, että tämän kurssin moduulit ovat pidempiä ja sisältävät enemmän labraharjoituksia kuin ensimmäinen kurssi. Kurssin pääpaino oli Pythonissa ja huomattavasti eniten aikaa meni Python-labraharjoituksiin. Harjoituksia oli paljon, mutta niistä oppi enemmän kuin tekstin lukemisesta, monivalintatesteistä tai opetusvideoista. Osa labratehtävistä oli haastavia, mutta hetken mietittyä ja koodia pyöriteltyäni ne kuitenkin onnistuivat. Harjoituksissa on myös vinkkejä tehtävien suorittamiseen, joita tuli silloin tällöin hyödynnettyä. Jokaisen harjoituksen jälkeen voi myös tarkistaa oikeat vastaukset seuraavasta moduulin osiosta.

Kurssilla tutustuttiin myös NumPy- ja pandas-kirjastoihin. Muitakin kirjastoja mainittiin, mutta tässä sertifikaatissa ei niihin perehdytä. Varsinkin pandas-kirjasto vaikutti hyödylliseltä data-analysoinnissa. Viimeisen moduulin projektiharjoituksessa pääsi kunnolla tutkimaan dataa, jonka oli itse luonut Python-funktioilla.

4.3 Kurssi 3: Go Beyond the Numbers: Translate Data into Insights

Moduuli 1.

Aloitan uuden kurssin alustalla, ja kuten aiemmissa ensimmäisissä moduuleissa, kurssin sisältö esitellään lyhyesti. Moduuli on melko lyhyt ja koostuu muutamasta videotallenteesta, lukemisesta sekä monivalintakokeesta. Alan ammattilaiset kertovat myös lyhyitä tarinoita aiheeseen liittyen, mikä tuo sisältöön käytännön näkökulmaa.

Moduulissa tutustutaan myös tutkivaan data-analyysiin (EDA, Exploratory Data Analysis), jolla tarkoitetaan aineiston alustavaa tarkastelua ja analysointia ennen varsinaista mallinnusta. EDA:n tavoitteena on ymmärtää datan rakennetta, havaita poikkeamia ja tunnistaa mahdollisia muuttujien välisiä yhteyksiä. Monivalintakokeen kysymykset ovat helppoja, ja läpäisen kokeen ensimmäisellä yrityksellä saaden 97,5 % pisteistä.

Moduuli 2.

Toinen moduuli on huomattavasti pidempi kuin ensimmäinen. Eniten aikaa kuluu kahteen labraharjoitukseen, joissa on todella paljon tehtäviä. Ensimmäisessä harjoituksessa tutkitaan data-aineiston sisältöä pandas-kirjaston funktioiden avulla sekä tutustutaan myös matplotlib-kirjastoon.

Toisen harjoituksen aiheena on tutkivan data-analyysin (EDA) jäsentäminen Pythonilla. Moduulissa on myös useita opetusvideoita, mutta tällä kertaa ne ovat huomattavasti pidempiä, jopa yli 10 minuutin pituisia. Videoissa esitellään todellisia työelämän tilanteita, joissa Pythonia hyödynnetään erilaisten analytiikkaan liittyvien ongelmien ratkaisemisessa. Saan monivalintakokeen ensimmäisellä yrityksellä 77,5 % pisteistä ja toisella yrityksellä täydet 100 %.

Moduuli 3.

Kolmas moduuli koostuu parista labraharjoituksesta, useasta pidemmästä videotallenteesta, lukumateriaalista sekä monivalintakokeesta. Harjoituksissa

jatketaan samalla teemalla, mutta tällä kertaa aiheena on datan validointi ja puhdistaminen Pythonilla.

Puuttuvaa dataa käsitellään eri kirjastojen, kuten NumPyn, pandasin, matplotlibin, Plotlyn ja seabornin avulla. Videoiden aiheina ovat muun muassa poikkeamien havaitseminen sekä syötteiden validointi Pythonilla. Monivalintakoe menee ensimmäisellä yrityksellä läpi pistein 100 %.

Moduuli 4.

Neljännessä moduulissa aiheena on datan visualisointi ja esittäminen. Moduuli poikkeaa kurssin aiemmista moduuleista, sillä siinä ei ole yhtään labraharjoitusta. Pythonin sijaan tutustutaan Tableau Public -ohjelmaan kahden harjoituksen kautta. Tableau on datavisualisointityökalu, jota käytetään interaktiivisten kuvaajien ja koontinäyttöjen luomiseen sekä analyysitulosten havainnolliseen esittämiseen.

Tableau on minulle entuudestaan tuttu, sillä sitä käytettiin Google Data Analytics -sertifikaatin aikana. Ensimmäisessä harjoituksessa muokataan Tableaun avulla data-aineistoa, josta laaditaan visualisointeja. Toisessa harjoituksessa Tableauta käytetään interaktiivisen koontinäytön suunnitteluun. Monivalintakoe menee jälleen ensimmäisellä kerralla läpi pistein 100 %.

Moduuli 5.

Kurssin viimeisessä moduulissa jatketaan Automatidata-projektin työstämistä. Projektissa suoritan tutkivan data-analyysin (EDA) projektiin liittyvälle aineistolle. Lisäksi käytän Tableuata visualisointien laatimiseen johdon yhteenvedon tueksi, jotta sidosryhmät voivat paremmin tarkastella ja hyödyntää aineistoa. Pythonin avulla aineistosta muodostetaan laatikkojana sekä histogrammi. Tämä tehdään Jupyter Notebook -harjoituksena.

Kurssin yhteenveto

Sertifikaatin kolmas kurssi vei enemmän aikaa kuin aiemmat, sillä siihen kuuluvat labraharjoitukset olivat huomattavasti pidempiä. Harjoitukset suoritettiin tuttuun tapaan Jupyter Notebook -ympäristössä Pythonilla, joka on siinä mielessä

kätevää, ettei omalle tietokoneelle tarvitse asentaa mitään ylimääräistä. Huomasin, että kurssilla olevat videotallenteet olivat pidempiä kuin aikaisempien kurssien, koska opeteltavaa on paljon. Välillä tuntui, etten oikein sisäistänyt videoiden materiaalia.

Tällä kurssilla alettiin myös visualisoida dataa. Visualisoinnissa käytettiin työkaluna Tableau Publicia, jota olen käyttänyt kerran aiemmin toisessa sertifikaatissa. Työkalua hyödynnettiin parissa harjoituksessa, mutta sen opettelu oli ajoittain hankalaa, mikä johtui todennäköisesti laajasta ja monimutkaisesta käyttöliittymästä. Sain kuitenkin tehtyä harjoitukset lopulta valmiiksi.

4.4 Kurssi 4: The Power of Statistics

Moduuli 1.

Neljännen kurssin ensimmäinen moduuli alkaa tuttuun tapaan sen sisällön esittelyllä, jonka aiheena on tilastotiede Pythonilla. Huomaan heti, että tässä kurssissa on yksi moduuli enemmän kuin aiemmissa, mutta moduulien pituus on hieman lyhyempi ja labraharjoituksia on vähemmän. Videotallenteita on useita, ja niissä esitellään uusia data-analytiikan termejä, kuten A/B-testaus, keski-, hajonta- ja sijaintiluvut. Videot tuntuvat edelleen miellyttävämmiltä kuin pelkkä ruudulta lukeminen.

Moduulin labraharjoituksessa asetutaan data-analyytikon rooliin ja tutkitaan eri osavaltioiden ilmanlaatua tietoaineiston avulla hyödyntämällä NumPy- ja pandas-kirjastoja. Läpäisen monivalintakokeen ensimmäisellä yrityksellä saaden 85 % vastauksista oikein.

Moduuli 2.

Toinen moduuli keskittyy todennäköisyyslaskentaan. Moduuli sisältää useita videotallenteita, joissa käsitellään todennäköisyyslaskennan eri teorioita ja sääntöjä. Sisällössä selitetään myös eri todennäköisyysjakaumia ja sitä, miten niitä hyödynnetään tiedon analysoinnissa sekä tietoaineiston ymmärtämisessä.

Videoiden jälkeen on muutaman kysymyksen monivalintatestejä, joilla testataan, onko opiskelija sisäistänyt materiaalin.

Moduulin labraharjoituksessa selvitetään, mikä todennäköisyysjakauma sopii aineistoon, ja lasketaan Z-arvot tietoaineistosta poikkeamien havaitsemiseksi. Labraharjoituksesta jää erityisesti mieleen se, kuinka poikkeavat ääriarvot voivat vääristää keskiarvoa, keskihajontaa ja koko analyysin tuloksia. Monivalintakokeen kysymykset ovat helppoja, joten saan ensimmäisellä yrityksellä täydet 100 % vastauksista oikein.

Moduuli 3.

Kurssin kolmannen moduulin aiheena on otanta. Moduuli koostuu useista videotallenteista, luettavasta materiaalista, labraharjoituksesta sekä monivalintakokeesta. Videoissa käsitellään otantaprosessin eri vaiheita, todennäköisyysmenetelmiä ja otantaharhan vaikutusta aineistoon. Videoissa esitellään myös esimerkkitilanteita siitä, miten data-alan ammattilaiset hyödyntävät otostunnuslukuja perusjoukon parametrien arvioinnissa.

Aloitan moduulin labraharjoituksen, jonka tehtävänä on toteuttaa aineiston otanta, jotta sen analysointi olisi helpompaa. Tämä tehdään tuttuun tapaan Jupyter Notebook -ympäristössä Pythonilla hyödyntäen eri kirjastoja. Moduulin monivalintakoe menee ensimmäisellä yrityksellä läpi, ja saan 85 % pisteistä.

Moduuli 4.

Neljännessä moduulissa perehdytään opetusvideoiden kautta luottamusväleihin, niiden tulkintaan sekä muodostamiseen. Moduuli on melko lyhyt, sillä se sisältää vain muutaman noin kuuden minuutin mittaisen videotallenteen, pari luettavaa materiaalia, noin tunnin kestävän labraharjoituksen sekä lopuksi monivalintakokeen. Labraharjoituksessa tarkastellaan luottamusvälejä Pythonilla. Läpäisen monivalintakokeen ensimmäisellä yrityksellä juuri ja juuri saaden 80 % pisteistä.

Moduuli 5.

Kurssin viidennessä moduulissa tutustutaan hypoteesitestaukseen. Aihetta käsitellään videotallenteiden, lukumateriaalin sekä labraharjoituksen avulla. Materiaalissa perehdytään myös eri testeihin, kuten yhden- ja kahden otoksen t-testeihin sekä A/B-testaukseen. Moduulin labraharjoituksessa tarkastellaan hypoteesitestausta Pythonilla. Monivalintakoe menee läpi ensimmäisellä yrityksellä saamalla 90 % kokonaispistemäärästä.

Moduuli 6.

Kurssin viimeisessä moduulissa jatketaan projektia, ja valitsen jälleen Automatidatan. Projektissa hyödynnetään aiemmin kurssilla käsiteltyjä tilastotieteen menetelmiä. Projekti etenee labraharjoituksella, jossa tutkitaan tietoaineistoa, toteutetaan hypoteesitesti ja lopuksi havainnot viestitään sidosryhmille. Projektin aikana jatketaan PACE-mallidokumentin täyttämistä. Kurssin päätösprojektin harjoitukset ovat erinomaisia, koska niissä päästään tekemään data-analyytikon työtä käytännössä. Tämän kurssin projektiosuus on myös huomattavasti lyhyempi kuin aiempien kurssien.

Kurssin yhteenveto.

Kurssissa oli tällä kertaa kuusi moduulia viiden sijaan. Ensimmäiset viisi moduulia noudattivat samaa rakennetta: ne alkoivat videotallenteilla ja lukumateriaalilla, ja moduulin lopussa oli laajempi labraharjoitus sekä monivalintakoe. Rakenne oli toimiva, sillä se tarjosi sopivassa suhteessa teoriaa ja käytännön harjoittelua.

Kurssin sisällön pääpaino oli tilastotieteessä. Videotallenteilla korostettiin, että aihetta käsitellään lähtötasolta, jossa suorittajalla ei oleteta olevan aiempaa kokemusta. Näin myös itse koin, sillä minulla ei ole aiempaa laajaa kokemusta tilastotieteestä. Kurssilla tarkasteltiin, miten data-alan ammattilaiset hyödyntävät tilastollisia menetelmiä aineiston analysoinnissa ja tulkinnassa sidosryhmien tietoon perustuvan päätöksenteon tukemiseksi.

4.5 Kurssi 5: Regression Analysis: Simplify Complex Data Relationships

Moduuli 1.

Viidennen kurssin ensimmäinen moduuli alkaa perinteiseen tapaan kurssin sisällön esittelyllä. Moduuli on lyhyt, sillä se sisältää ainoastaan videotallenteita, lukumateriaalia sekä monivalintakokeen. Videoilla esitellään regressioanalyysin peruskäsitteitä, kuten lineaarinen ja logistinen regressio. Data-analytiikka-alan ammattilainen myös kertoo omia kokemuksiaan siitä, miten hän on käytännössä tuottanut regressiomallien avulla hyödynnettävää tietoa. Regressioanalyysin toimintaperiaate esitellään lisäksi suhteessa koko sertifikaatin aikana käytettyyn PACE-menetelmään ja sen eri vaiheisiin. Kurssin ensimmäinen monivalintakoe menee läpi onnistuneesti saamalla 90 % pisteistä.

Moduuli 2.

Kurssin toisen moduulin aloitettuani huomaan, että labraharjoitukset ovat tehneet paluun. Moduuliin sisältyy kaksi noin tunnin mittaista harjoitusta ja tietysti videotallenteita, lukumateriaalia sekä monivalintakoe. Aloitan moduulin ensimmäisen labraharjoituksen, jonka aiheena on yksinkertaisen lineaarisen regression suorittaminen Pythonilla. Harjoitus suoritetaan tutussa Jupyter Notebook -ympäristössä, jossa osallistun fiktiiviseen data-analytiikkatiimin jäsenen rooliin.

Harjoituksen selkeät tehtävänannot, annetut vinkit sekä sertifikaatin aikana kehittyneet Python-taidot mahdollistavat harjoituksesta suoriutumisen ilman suurempia ongelmia. Moduulin toinen harjoitus jatkaa yksinkertaisen lineaarisen regression teemaa, mutta tällä kertaa tavoitteena on tulkita regressioanalyysin tuloksia Python-kirjastojen avulla. Läpäisen monivalintakokeen ensimmäisellä yrityksellä saaden 97,5 % pisteistä.

Moduuli 3.

Kolmannen moduulin rakenne on sama kuin toisen, mutta kahden harjoituksen sijaan harjoituksia on vain yksi. Teemana on monimuuttujainen lineaarinen regressio ja sen tulkinta Pythonin avulla. Labraharjoituksessa suoritan useita

monimuuttujaisia lineaarisia regressioanalyysejä Pythonilla, joissa myyntiä arvioidaan useiden selittävien muuttujien yhdistelmien perusteella. Harjoituksessa vastataan myös erilaisiin kysymyksiin omien analyysien tulosten pohjalta. Oikeat vastaukset voi tarkistaa malliratkaisusta harjoituksen jälkeen. Teen vielä lopuksi monivalintakokeen, joka menee ensimmäisellä kerralla läpi pistein 82,5 %.

Moduuli 4.

Neljännessä moduulissa mennään syvemmälle viime kurssin viidennen moduulin aiheeseen eli hypoteesitestaukseen. Materiaalissa tutustutaan uusiin testimenetelmiin, kuten khiin neliö -testiin sekä erilaisiin ANOVA-testeihin. Khiin neliö -testillä tarkastellaan, onko kahden kategorisen muuttujan välillä tilastollisesti merkitsevää yhteyttä. ANOVA-testillä puolestaan vertaillaan useamman kuin kahden ryhmän keskiarvoja ja arvioidaan, poikkeavatko ne toisistaan tilastollisesti merkitsevästi.

Moduulin labraharjoituksessa käytetään aineistoa yksisuuntaisen ANOVA:n ja post hoc -testin suorittamiseen Pythonilla. Testien tulokset välitetään lyhyesti fiktiivisille sidosryhmille. Harjoituksessa hyödynnetään tuttuun tapaan Python-kirjastoja, jotka tekevät koodin tuottamisesta helpompaa, ja niistä saa kätevästi tulostettua valmiit kaaviot. Monivalintakoe menee jälleen kerran läpi ensimmäisellä kerralla, mutta tällä kertaa sain 90 % pisteistä.

Moduuli 5.

Moduuli koostuu yhdestä labraharjoituksesta, videotallenteista, lukumateriaalista sekä monivalintakokeesta. Moduulin aiheena on jo aiemmin tällä kurssilla mainittu logistinen regressio. Videomateriaalia aiheesta on runsaasti, ja se antaa selkeän kokonaiskuvan siitä, miten erilaisia logistisia regressiomalleja hyödynnetään analytiikassa. Aloitan seuraavaksi moduulin labraharjoituksen, jossa tehtävänä on suorittaa binominen logistinen regressio Pythonin avulla. Harjoitus sujuu hyvin ja saan siihen kuuluvat tehtävät tehtyä ja kysymykset vastattua. Läpäisen moduulin monivalintakokeen ensimmäisellä yrityksellä ja saan 90 % pisteistä.

Moduuli 6.

Aloitan kurssin viimeisen moduulin, joka on tyypilliseen tapaan projektin tekemistä. Kolmesta projektivaihtoehdosta valitsen taas minulle entuudestaan tutun Automatidatan. Projektiharjoitus tehdään tutussa Jupyter Notebook -ympäristössä, ja tehtävänä on laskea kuvailevat tilastot, laatia regressiomalli aineistosta sekä lopuksi laatia tiivis yhteenveto analyysien tuloksista Automatidatan datatiimille. Kurssin projektiosuus on tällä kertaa huomattavasti laajempi kuin aiempien kurssien projektiosuudet.

Kurssin yhteenveto.

Sertifikaatin viidennellä kurssilla teemana olivat regressioanalyysit. Kurssilla perehdyttiin monipuolisesti eri regressiomalleihin, testausmenetelmiin sekä mallien arviointiin ja tulkintaan. Kattavat Jupyter Notebook -ympäristössä suoritettavat labraharjoitukset olivat jälleen kurssin kohokohta oppimisen kannalta. Pidin erityisesti siitä, että harjoituksia ei ollut liikaa, vaan jokaisessa moduulissa oli yksi pidempi harjoitus moduulin aiheesta. Myös viime kurssista pitämäni moduulirakenne pysyi melko samanlaisena tässä kurssissa.

Lukumateriaalia oli tällä kertaa minulle sopivasti eli vähän, sillä opetusmateriaalin paino oli videotallenteissa. Kurssin videotallenteiden vetäjä oli erityisen hyvä tällä kurssilla. Videot olivat sopivan pituisia, monipuolisia ja laadukkaita. Monivalintakokeet tuntuivat erityisen helpoilta, koska kaikki menivät ensimmäisellä yrityksellä läpi.

4.6 Kurssi 6: The Nuts and Bolts of Machine Learning

Moduuli 1.

Kuudennen kurssin ensimmäinen moduuli aloitetaan jälleen kurssin aiheen esittelyllä, joka on tällä kertaa koneoppiminen. Kurssilla tutustutaan koneoppimisen eri tyyppeihin, niiden tarkoitukseen ja rakennetaan omia koneoppimismalleja. Alussa mainitaan myös, että tällä kurssilla ei perehdytä tekoälyyn, vaan keskitytään ohjattuun ja ohjaamattomaan koneoppimiseen.

Moduuli koostuu useista videotallenteista, lukumateriaalista sekä monivalintakokeesta. Ensimmäisessä moduulissa ei ole labraharjoitusta lainkaan, mikä on ymmärrettävää koneoppimisen laajuuden vuoksi. Videoista ja lukumateriaalista saa hyvän kokonaiskuvan koneoppimisesta yleisesti. Teen vielä lopuksi moduulin monivalintakokeen, ja se vaatii tällä kertaa kaksi yritystä. Ensimmäisellä yrityksellä saan 70 % pisteistä ja toisella yrityksellä läpäisen kokeen saamalla 85 % pisteistä.

Moduuli 2.

Heti toisen moduulin alussa mainitaan, että vihdoin päästään rakentamaan koneoppimismalleja alusta loppuun, mikä herättää mielenkiinnon aiheeseen. Huomaan moduulin sisällöstä, että siihen kuuluu kaksi noin tunnin mittaista labraharjoitusta. Muuta materiaalia ei ole kovin paljon, vaan ainoastaan muutama videotallenne ja lukumateriaali sekä monivalintakoe.

Seuraavaksi aloitan moduulin ensimmäisen labraharjoituksen, jossa tehtävänä on selvittää rakennettavaa mallia varten sopivat muuttujat suorittamalla piirteiden muokkausta aineistoon. Nämä havainnot hyödynnetään seuraavassa harjoituksessa, jossa rakennetaan ennustemalli. Aloitan moduulin toisen harjoituksen, jonka alussa aiemmassa harjoituksessa muokattu aineisto tuodaan ympäristöön. Harjoituksessa rakennetaan aineistosta Naive Bayes -malli Pythonin avulla hyödyntämällä pandas- ja scikit-learn-kirjastoja. Naive Bayes on todennäköisyyspohjainen luokittelumalli, jota käytetään erityisesti kategoristen lopputulosten ennustamiseen esimerkiksi tekstiluokittelussa.

Labraharjoitus antaa hyvän käsityksen siitä, kuinka rakennettua mallia voidaan käyttää ennakoinnissa, kuten tässä tapauksessa ennustettaessa kestääkö NBA-pelaajan ura yli viisi vuotta. Moduulin lopuksi teen vielä monivalintakokeen, josta saan ensimmäisellä yrityksellä 95 % pisteistä.

Moduuli 3.

Kolmannen moduulin aiheena on K-means-klusterointi. Moduuli koostuu lukumateriaalista, videotallenteista, labraharjoituksesta sekä monivalintakokeesta. Moduulin labraharjoituksessa asetutaan data-analyytikon

rooliin ja rakennetaan aineistosta K-means-malli. Harjoituksessa käytetään julkista tietoaineistoa, joten dataan suoritetaan ensin EDA-prosessi ennen mallin rakentamista.

Harjoituksen suorittamisessa on hieman ongelmia, mutta lopulta saan suurimman osan kohdista tehtyä oikein. Harjoituksen jälkeen tarkistan oikeat vastaukset mallivastauksesta. Lopuksi teen moduulin monivalintakokeen, jonka läpäisen ensimmäisellä yrityksellä saamalla 87,5 % pisteistä.

Moduuli 4.

Kurssin neljäs moduuli käsittelee koneoppimisen puupohjaista mallintamista. Moduuli on kurssin huomattavasti pisin, sillä se koostuu kolmesta harjoituksesta, useista videotallenteista, lukumateriaalista sekä monivalintakokeesta. Useiden videoiden jälkeen aloitan vihdoin moduulin ensimmäisen labraharjoituksen. Ensimmäisessä harjoituksessa tehtävänä on rakentaa päätöspuu Pythonilla. Oma päätöspuuni näyttää huomattavasti erilaiselta kuin mallivastaus, mutta yhtä arvoa muuttamalla saan sen korjattua.

Moduulin toisessa harjoituksessa rakennetaan Pythonilla satunnaismetsä-malli. Harjoitus tuottaa hieman ongelmia, mutta lopulta saan virheilmoitukset korjattua ja mallin valmiiksi. Viimeisessä harjoituksessa tehtävänä on rakentaa Pythonilla XGBoost-malli. XGBoost (Extreme Gradient Boosting) on koneoppimismenetelmä, joka perustuu useiden päätöspuiden yhdistämiseen ja jota käytetään ennustetarkkuuden parantamiseen erityisesti luokittelu- ja regressiotehtävissä. Harjoitus onnistuu ilman suurempia ongelmia. Läpäisen vielä monivalintakokeen ensimmäisellä yrityksellä saamalla 97,5 % pisteistä.

Moduuli 5.

Aloitan kurssin viimeisen moduulin, joka on jälleen kerran Automatidata-projektin työstämistä. Projektin aikana täytetään PACE-menetelmädokumenttia, jonka pohjalta vastataan myös Jupyter Notebook -ympäristön kysymyksiin. Tällä kertaa projektissa rakennetaan Pythonilla satunnaismetsä-malli, joka on jo tuttu aiemmasta moduulista. Laadin seuraavaksi tuloksista tiiviin yhteenvedon fiktiiviselle Automatidatan datatiimille. Projektiharjoituksen tekeminen onnistuu

ilman suurempia ongelmia. Muutama kysymys jää vastaamatta, mutta saan lopulta harjoituksen valmiiksi.

Kurssin yhteenveto.

Sertifikaatin kuudennen kurssin pääpaino oli koneoppimisessa. Koko kurssi käsitteli aihetta monipuolisesti, ja hyödyllisiä labraharjoituksia oli useita. Pääsin testaamaan useiden erilaisten koneoppimismallien rakentamista Jupyter Notebook -ympäristössä Pythonilla. Osa niistä tuotti ongelmia, mutta sain kuitenkin jokaiseen soluun koodia sekä vastaukset suurimpaan osaan kysymyksistä.

Moduuleja oli tällä kertaa viisi kuuden sijaan, mutta ainakin neljäs moduuli oli todella pitkä. Videotallenteet olivat edelleen hyvälaatuisia ja sopivan pituisia. Lukumateriaalia oli paljon, ja osa teksteistä oli pidempiä kuin yleensä.

4.7 Kurssi 7: Google Advanced Data Analytics Capstone

Sertifikaatin viimeinen kurssi on vaihtoehtoisen projektin tekeminen. Projektia ei ole pakko suorittaa sertifikaatin läpäisemiseksi, mutta päätän kuitenkin tehdä sen. Projekti on hyvin samanlainen kuin aiempien kurssien projektit, mutta tällä kertaa siinä hyödynnetään koko sertifikaatin aikana opittuja taitoja.

Projektin tehtävänä on analysoida fiktiivisen Salifort Motors -yrityksen henkilöstökyselyn aineistoa ja rakentaa sen pohjalta ennustemalli, jonka tavoitteena on parantaa henkilöstöretentiota. Projektissa saa itse päättää, käyttääkö regressiomallia vai puupohjaista koneoppimismallia, ja valitsen näistä regressiomallin.

Projektin työstäminen alkaa tuttuun tapaan PACE-menetelmädokumentin laatimisella. Seuraavaksi käynnistän projektin Jupyter Notebook -harjoituksen. Tuon ympäristöön tarvittavat kirjastot ja puhdistan aineiston. Muutaman vaiheen jälkeen siirrytään datan visualisointiin. Tämä toteutetaan laatimalla aineistosta muutama laatikkojanakuvio ja lukumääräkaavio Pythonilla. Lopuksi rakennetaan regressiomalli.

Projektin yhteenveto.

Sain projektin valmiiksi ilman suurempia ongelmia, sillä hyödynsin apuna aiempia harjoituksia, erityisesti regressiomalliharjoitusta. Projekti oli myös hyvin samanlainen kuin aiemmat sertifikaatin projektit, mikä helpotti sen toteuttamista huomattavasti. Kaikkiin kysymyksiin en välttämättä saanut järkevää vastausta, mutta sain regressiomallin rakennettua. En tehnyt aiemmassa sertifikaatissa capstone-projektiosuutta, mutta onneksi tein sen tällä kertaa, sillä se toimi hyvänä kertauksena koko sertifikaatin aikana opittuun sisältöön. Pidin myös siitä, että projektissa sai itse valita menetelmät ja mallin. Lopuksi tein projektin arvioinnin, jossa vastattiin myös siihen, suorittiko projektin vai ei. Tämän jälkeen sain sähköpostiini onnitteluviestin siitä, että sertifikaatti on suoritettu.

5 POHDINTA

Opinnäytetyön tarkoitus oli suorittaa Courseran Google Advanced Data Analytics -sertifikaatti. Sertifikaatin suorittamisesta laadittiin päiväkirja, jossa kuvataan etenemistä, kurssien sisältöä sekä sertifikaatin hyötyä opiskelijalle. Päiväkirjaan tehtiin merkinnät jokaisen kurssin moduulin keskeisestä sisällöstä sekä kurssikohtaiset yhteenvedot.

Koin yleisesti kurssin suorittamisen hyödylliseksi ja mielenkiintoiseksi. Kurssien sisältö oli monipuolista, laadukasta ja opettavaista. Opetusvideoita oli runsaasti, ne esittivät käsiteltävät asiat selkeästi ja tiiviisti. Lukumateriaali toimi pääosin kertauksena, ja osa materiaalista oli ajoittain turhan pitkiä. Moduuleihin sisältyneet lyhyet monivalintatestit tukivat oppimista ja mahdollistivat oman osaamisen arvioinnin. Pisteytetyistä kokeista suoriuduin hyvin, sillä sain suurimman osan kokeista ensimmäisellä yrityksellä läpi. Kysymykset olivat sopivan haastavia ja edellyttivät perehtymistä kurssimateriaaliin.

Laboratorioharjoitukset muodostivat sertifikaatin keskeisimmän ja mielekkäimmän osa-alueen. Laboratoriot sisälsivät käytännönläheisiä tehtäviä, jotka kehittivät Python-ohjelmointikielen ja sen kirjastojen hyödyntämistä data-analytiikan ongelmanratkaisussa. Projektiharjoitusten aikana laadittiin myös PACE-menetelmädokumenttia, mikä auttoi jäsentämään analyysiprosessia systemaattisesti ja kehitti kykyäni perustella analyysiin liittyviä ratkaisuja ammatillisesta näkökulmasta.

Tietoperustan laatiminen, lähteiden etsiminen sekä niihin perehtyminen toimivat osittain kertauksena sertifikaatin aikana käsitellyille aiheille. Hyvien lähteiden löytäminen osoittautui välillä haastavaksi, sillä monet tieteelliset julkaisut olivat maksumuurin takana. Vaikka tietoperusta jäi laajuudeltaan melko tiiviiksi, se kattaa rajauksen mukaisesti työn keskeiset käsitteet ja teemat.

Sertifikaatin suorittaminen vahvisti valmiuksiani toimia data-analytiikan parissa työelämässä. Kurssin aikana kehittyivät erityisesti analyyttinen ajattelu, Python-ohjelmointiosaaminen sekä kyky jäsentää ja dokumentoida analyysiprosessi

systemaattisesti. Nämä taidot ovat keskeisiä nykyaikaisessa data-analytiikassa, jossa teknisen osaamisen lisäksi korostuvat ymmärrettävä raportointi ja perusteltu päätöksenteko. Sertifikaatti tarjoaa myös konkreettisen näytön osaamisesta, mikä voi parantaa kilpailukykyä työmarkkinoilla ja tukea siirtymistä data-analytiikkaan liittyviin työtehtäviin tulevaisuudessa.

Kokonaisuutena opinnäytetyö ei ainoastaan syventänyt teknistä osaamistani, vaan vahvisti myös ammatillista identiteettiäni data-analytiikan osaajana.

LÄHTEET

Almaci, A. 17.12.2023. NumPy for Data Analysis. Luettavissa: https://medium.com/@aysealmaci/numpy-for-data-analysis-dd52e5635d5b Luettu: 21.2.2026.

Carreiro, J. 15.7.2023. The Role of Machine Learning in Data Analysis: Key Benefits and… Luettavissa: https://www.ironhack.com/us/blog/the-role-of-machine-learning-in-data-analysis Luettu: 22.2.2026.

Clark, B. s.a. What is Scikit-Learn (Sklearn)? Luettavissa: https://www.ibm.com/think/topics/scikit-learn Luettu: 22.2.2026.

Clement, M. 8.10.2024. Python 101: Introduction to Python as a Data Analytics Tool. Luettavissa: https://dev.to/clement_mwai/python-101-introduction-to-python-as-a-data-analytics-tool-nef Luettu: 19.2.2026.

Cote, C. 14.12.2021. What Is Regression Analysis in Business Analytics? Luettavissa: https://online.hbs.edu/blog/post/what-is-regression-analysis Luettu: 22.2.2026.

Coursera Staff. 13.11.2025. What is Pandas Python Library? Luettavissa: https://www.coursera.org/articles/what-is-pandas-python Luettu: 22.2.2026.

Coursera Staff. 4.6.2025. Model Evaluation: Assessing the Performance of Machine Learning Models. Luettavissa: https://www.coursera.org/articles/model-evaluation Luettu: 22.2.2026.

Day, F. 15.7.2025. Why Every Data Scientist Should Know Matplotlib. Luettavissa: https://www.nobledesktop.com/blog/why-learn-matplotlib-for-data-science Luettu: 22.2.2026.

Hayriye, A. 3.11.2024. The Heart of Python: History, Clean Code and the Zen of Programming. Luettavissa: https://anill-hayriye.medium.com/the-heart-of-python-history-clean-code-and-the-zen-of-programming-92db86a56847 Luettu: 20.2.2026.

Karl, T. 31.5.2024. Benefits of Python for Data Analytics Explained. Luettavissa: https://www.newhorizons.com/resources/blog/benefits-of-python-for-data-analytics Luettu: 21.2.2026.

McKinney, W. 2022. Python for Data Analysis. 3rd ed. Sebastopol: O'Reilly Media. Luettavissa: https://wesmckinney.com/book/ Luettu: 22.2.2026.

NumPy s.a. What is NumPy? Luettavissa: https://numpy.org/doc/stable/user/whatisnumpy.html Luettu: 21.2.2026.

Python. Python events calendar 2026. Luettavissa: https://www.python.org/events/python-events/ Luettu: 20.2.2026.

Saari, A & Ahvonen, J. 20.1.2026. Millä sitä tänään koodaisi? Katsaus ohjelmistokielten maailmaan. Luettavissa: http://urn.fi/URN:NBN:fi-fe202601205114 Luettu: 20.2.2026.

Sarker, IH. 12.7.2021. Data Science and Analytics: An Overview from Data-Driven Smart Computing, Decision-Making and Applications Perspective. Luettavissa: https://pmc.ncbi.nlm.nih.gov/articles/PMC8274472/ Luettu: 18.2.2026.

Scikit-learn s.a. Luettavissa: https://scikit-learn.org/stable/getting_started.html Luettu: 22.2.2026.

Seaborn s.a. An Introduction to seaborn. Luettavissa: https://seaborn.pydata.org/tutorial/introduction Luettu: 22.2.2026.

Thavali, A. 16.10.2025. Understanding the Data Analysis Process. Luettavissa: https://medium.com/@ankush.thavali/understanding-the-data-analysis-process-74570d6c6ca2 Luettu: 19.2.2026.

Unwin, A. 31.1.2020. Why Is Data Visualization Important? What Is Important in Data Visualization? Luettavissa: https://doi.org/10.1162/99608f92.8ae4d525 Luettu: 22.2.2026.

Verbina, E. 24.9.2025. Why Is Python So Popular In 2025? Luettavissa: https://blog.jetbrains.com/pycharm/2025/09/why-is-python-so-popular/ Luettu: 20.2.2026.

Walker, J. 2019. Hypothesis tests. BJA Education, s. 227–231. Luettavissa: https://doi.org/10.1016/j.bjae.2019.03.006 Luettu: 22.2.2026.

World Economic Forum. The Future of Jobs Report 2025. Luettavissa: https://www.weforum.org/publications/the-future-of-jobs-report-2025/ Luettu: 19.2.2026.