VIININ LAADUN ANALYYSI

Lappeenrannan-Lahden teknillinen yliopisto LUT

Laskennallisen tekniikan työkurssi, Harjoitustyö

2022

Ilmari Vahteristo, Arno Törö, Santeri Kokkonen

Tarkastaja: Matylda Jablonska-Sabuka

TIIVISTELMÄ

Lappeenrannan-Lahden teknillinen yliopisto LUT
School of Engineering Science
Laskennallinen tekniikka

Ilmari Vahteristo, Arno Törö, Santeri Kokkonen

Viinin laadun analyysi

Harjoitustyö
2022
31 sivua, 14 kuvaa, 9 taulukkoa
Tarkastaja: Matylda Jablonska-Sabuka

Hakusanat: viini, laatu, mallinnus, regressioanalyysi, epätasapainoinen data

Tässä työssä analysoitiin viinin fysikokemiallisten ominaisuuksien vaikutusta portugalilaisten *Vinho Verde* viinien laatuun. Datasettinä käytimme vuonna 2009 kerättyä puna- ja valkoviini dataa. [1] Viinin laadun tutkiminen on tärkeää, sillä viinin laatuun vaikuttaviin muuttujiin voidaan vaikuttaa hyvin vahvasti tuotantovaiheessa.

Analyysissä tutkittiin lineaarisen- ja logistisen regression avulla yksittäisten muuttujien vaikutusta viinien laatuun, sekä sovitettiin regressioon perustuvia ennustusmalleja. Mallien sopivuutta arvioitiin regressio sekä luokitteleville malleille suunnatuilla mittareilla.

Viinin laadun mallintamiseen käytettiin lineaarista mallia, neuroverkkoja ja satunnaismetsiä, joita optimoitiin dataa prosessoimalla ja hyperparametrien optimoinnilla. Molempia menetelmiä suoritettiin automatisoidusti, että manuaalisesti käsin. Jokaiselle mallille etsittiin kaksi hyvää tapausta, missä toisessa painotettiin tarkkutta ja toisessa korkeaa f1-arvoa.

Yksittäisistä tekijöistä eniten viinin laatuun vaikuttavat alkoholi ja etikkahappo. Parhaat mallit molemmille viineille saatiin satunnaismetsän avulla, joissa korkein tarkkuus saavutettiin ilman datan oversamplausta. Korkein f1 arvo saavutettiin yli- ja/tai alinäytteistämällä SMOGN menetelmällä tai manuaalisesti valitsemalla laadun arvot, joiden mittaukset ylinäytteistettiin monta kertaa treenausdataan.

SYMBOLI- JA LYHENNELUETTELO

MAE	Mean Absolute Error
MSE	Mean Squared Error
SMOGN	Synthetic Minority Over-sampling Technique for regression
RSS	Residual Sum of Squares

SISÄLLYSLUETTELO

1 JOHDANTO

Viinit sisältävät monia objektiivisesti mitattavia fysikokemiallisia ominaisuuksia, kuten tiheys, alkoholipitoisuus ja pH arvo. Mikäli näiden tekijöiden vaikutus viinin laatuun (viiniasiantuntijoiden keskimääräinen arvosana) tiedetään, pystytään kasvatusmenetelmillä (mm. käytetty lannoite) sekä kypsytystavoilla (mm. kypsytysaika, lisäaineet) yrittää tuottaa mahdollisimman laadukasta viiniä.

Työmme tavoitteena on arvioida fysikokemiallisten tekijöiden vaikutusta viinin laatuun erilaisilla regressioanalyysin menetelmillä, ja antaa viinin valmistajille ideoita, kuinka parantaa viinin laatua. Tavoitteenamme on myös luoda koneoppimismalli viinin laadun arvioimiseksi fysikokemiallisista mittauksista, joka mahdollistaisi viinin laadun tarkkailun mm. kypsyttämisvaiheessa avaamatta säiliöitä. Lineaarinen ja logistinen regressio, sekä viinin laadun ennustus ovat erillisiä tutkimusmenetelmiä, joiden tuloksia vertaamme toisiinsa ja yhdistämme päätelmät.

1.1 Tausta

Viini on suuressa osassa maailmaa käytetty alkoholi tuote, jonka markkinaosuus oli vuonna 2022 yli US $510 miljardia. [2]

Missä raha siellä tutkimukset. Viineistä on tehty monia tutkimuksia kuten eri tekijöiden vaikutus kuluttajahintaan [3], viinin laadun mallinnus fysikokemiallisten tekijöiden avulla [1], sekä sään vaikutus viinin laatuun [4].

Tutkimuksemme tavoitteena on arvioida yksittäisen fysikokemiallisen tekijän vaikutus viinin laatuun valmistusprosessien optimoimiseksi ja tuottaa viinin laadun ennustava koneoppimismalli, jota voisi hyödyntää mm. valvomalla viinin fysikokemiallisia ominaisuuksia. Tuloksia hyöndyntämällä voidaan tuottaa jatkuva arvio tai ennuste viinin laadusta.

1.2 Tutkimusmetodologia

Käytämme puna- ja valkoviinille samoja metodeja erikseen. Emme sovita malleja käyttäen puna-/valkoviini kategorista muuttujaa viinien erilaisuuden takia. Aloitamme tutkimuksen tutkimalla dataa ja arvioimalla sen laatua. Kun data on tarvittaessa putsattu ja käyttökelpoisuus selvitetty, sovitamme dataan lineaarisen mallin arvioidaksemme yksittäisten muuttujien vaikutusta viinin laatuun. Välttyäksemme multikorrelaatiolta, tutkimme datasta luotua korrelaatiomatriisia, ja poistamme keskenään korreloivista selittävistä muuttujista sen, joka korreloi vähemmän ennustettavan muuttujan (laatu) kanssa.

2 METODIEN ESITTELY

2.1 Lineaarinen regressio

Lineaarinen regressiomalli kuvaa selitettävän muuttujan y ja selittävien muuttujien $x_1, x_2, x_3, \ldots, x_p$ välistä yhteyttä. Yhtälö on muotoa

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon, \tag{1}$$

missä $\beta_0, \beta_1, \beta_2, \beta_p$ ovat regressiokertoimia ja ε virhetermi.

Tässä työssä regressiokertoimet estimoidaan minimoimalla residuaalien neliöiden summa

$$RSS = \sum_{i=1}^{n} (y_i - \sum_{j=0}^{p} \beta_j x_{ij})^2, \tag{2}$$

missä n on rivien määrä.

2.2 Logistinen regressio

Logistista regressiota käytetään, kun selitettävänä muuttujana toimii kaksiluokkainen muuttuja joka saa arvon 0 tai 1. Tässä analyysissä käytetään ordinaalista logistista regressiota (engl. ordinal logistic regression), joka on logistisen regression jatke [5]. Ordinaalinen log-malli eroaa normaalista siten, että selitettävä muuttuja voi saada moniluokkaisia arvoja joilla on järjestysasteikko.

Logistisen mallin kertoimet estimoidaan suurimman uskottavuuden menetelmällä, mikä käytännössä tarkoittaa että menetelmä tuottaa mallin kertoimille arvot, jotka maksimoivat todennäköisyyden mallille saada aineistosta havaitut arvot. Regressiosta saadut tulokset ovat logaritmisia todennäköisyyksien suhteita tapahtumalle, tässä analyysissä siis todennäköisyys viinille tiettyyn laatuluokkaan kuulumiselle.

Käytetään logaritmisille todennäköisyyksien suhteille merkintää $logit$ mikä on johdettu seuraavasti

$$\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = logit(P(Y \leq j)), \tag{3}$$

missä j on selitettävän muuttujan luokkanumero. Logistinen malli voidaan esittää seuraavasti

$$logit(P(Y \leq j)) = \beta_{j0} - (\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p), \tag{4}$$

missä j on selitettävän muuttujan luokkanumero ja p on selittävien muuttujien määrä.

2.3 Neuroverkko

Neuroverkko koostuu neuroneista, synapseista, painoista ja funktioista. Neuroverkkoja voidaan käyttää sekä luokittelu-, että regressio-ongelmissa. Neuroverkot toimivat hyvin kompleksisten, ja epälineaaristen funktioiden mallinuksessa.

Yksinkertaisuudessaan gradienttin perustuva neuroverkko ottaa sisään numeerisen arvon, joka kulkeutuu verkon ja sen painojen läpi. Viimeisen kerroksen jälkeen neuroverkko antaa numeerisen arvon. Tälle numeeriselle arvolle lasketaan häviö valitulla Loss funktiolla vertaamalla oikeaa vastausta sekä neuroverkon arviota. Tämän jälkeen häviölle lasketaan gradientti jokaisen painon suhteen, ja painoja muutetaan gradientin mukaan.

Erilaisia neuroverkkoja on todella paljon. Tässä työssä käytämme sekventiaalista neuroverkkoa, jossa jokaisella kerroksella on ainoastaan yksi sisääntulo tensori, ja yksi ulostulo tensori. Neuroverkon arkkitehtuuria on etsitty kokeilemalla järjestelmällisesti eri arkkitehtuureja *käsittelemättömällä* datalla. Käsitellylle datalle arkkitehtuuria ei ole etsitty uudestaan hyperparametrien optimoinnin kompleksisuuden vuoksi.

2.4 Satunnaismetsä

Satunnaismetsä (engl. Random Forest) on algoritmi, joka luo monta päätöspuuta satunnaisesti valituilla datan arvoilla. Satunnaismetsä valittiin päätöspuun sijasta, sillä satunnaismetsä on yleisemmin käytetty, eikä se ylisovita yhtä helposti. Yksittäinen päätöspuu sisältää monia sisäkkäisiä if-else lauseita. Raja-arvot lauseille etsitään vertaamalla parametrin arvoa, ja valitsemalla arvot, jotka minimoivat käytetyn häviön funktion. Satunnaismetsän puut toimivat toisistaan erillisinä ja jokainen niistä tuottaa oman ennustuksen. Tämän jälkeen kaikkien puiden ennustuksista otetaan keskiarvo, joka on satunnaismetsän ennustus.

2.5 Datan prosessointi

Tässä osiossa esitellään lyhyesti työssä käytettyjä datan esiprosessointimenetelmiä. Menetelmiä ei olla käytetty jokaisessa kohdassa. Seuraavissa kaavoissa x_i kuvaa muuttujan x yksittäistä mittausta, ja $\boldsymbol{x}$ kuvaa muuttujan x kaikkia mittauksia.

Datan prosessointi tehtiin ennustusmalleissa koko datasetille, eikä erikseen testi- ja treenidatalle. Tämä johtaa epäsuoraan, ja todennäköisesti erittäin pieneen informaation vuotoon, sillä testi data on transformoitu samoilla parametreillä kuin treenidata. Lisää tästä tyypillisestä virheestä voi lukea lukea lähteestä [6]. Tämä virhe huomattiin liian myöhään, eikä ennustusmallien tuloksia ehditty korjata.

2.5.1 Normalisointi

Datan normalisoinnilla vältytään monilta numeerisilta ongelmilta, ja parannetaan mallien konvergoitumista. Mikäli dataa ei ole normalisoitu, mm. neuroverkko painottaa suuria muuttujia suhteettoman paljon, joka on usein epätoivottua. Data normalisoitiin jokaiseen ennustusmalliin jokaisen muuttujan kohdalla seuraavalla menetelmällä

$$x_{i_n} = \frac{x_i}{\max(\boldsymbol{x})}, \tag{5}$$

jossa x_{i_n} on normalisoitu muuttujan arvo.

2.5.2 Standardisointi

Datan standardisointi muokkaa mittausten jakaumaa siten, että mittausten keskiarvoksi tulee 0 ja hajonnaksi 1. Datan standardisointi on tärkeä menetelmä koneoppimismalleille ja tässä tutkimusessa käytimme datan standardisointiin seuraavaa menetelmää

$$x_{i_z} = \frac{x_i - \bar{x}}{\sigma}, \tag{6}$$

jossa x_{i_z} on standardisoitu arvo, $\bar{x}$ on $\boldsymbol{x}$ keskiarvo, ja σ on $\boldsymbol{x}$ keskihajonta.

2.5.3 Box-cox muunnos

Box-cox muunnos muuttaa epänormaalin selittävän muuttujan datan normaalijakautuneeksi, mikä mahdollistaa mm. laajempien tilastollisten testien suorittamisen. Box-cox muunnos tehdään seuraavalla menetelmällä

$$x_{i_B} = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & , \text{if } \lambda \neq 0 \\ \log x_i & , \text{if } \lambda = 0 \end{cases}, \tag{7}$$

jossa λ on etsitty arvo jota käyttämällä $\boldsymbol{x}$ jakauma on lähimpänä normaalijakaumaa.

Box-cox muunnos toimii ainoastaan positiivisilla arvoilla, joten tätä muutosta varten datasta poistettiin kaikki rivit, joissa joku mittauksista oli 0. (Ei ollut montaa)

2.5.4 Manuaalinen ylinäytteistäminen

Tässä analyysissa hyvät arvot ylinäytteistämiselle etsittin erikseen puna- ja valkoviinille manuaalisen kokeilun avulla. 'Monistus' kertoo, kuinka monta kertaa vastaavaa laatua sisältävät mittaukset kopioitiin testidatassa. Parhaat tulokset saavutettiin taulukoiden 1 ja 2 mukaan.

Taulukko 1. Punaviinin monistus taulukko.

Laatu	Monistus
3	8
4	2
5	1
6	1
7	1
8	7

Taulukko 2. Valkoviinin monistus taulukko.

Laatu	Monistus
3	6
4	1
5	1
6	1
7	1
8	2
9	10

2.5.5 SMOGN algoritmi

SMOGN algoritmi on algoritmi, joka generoi synteettisesti uusia mittauksia harvinaisille mittauksille, sekä käyttää satunnaista alinäytteistämistä yleisille mittauksille. SMOGN hyödyntää kahta yleisesti käytettyä ylinäytteistämismetodia regressiolle: SMOTER [7], sekä kohinan lisääminen. Algoritmi päättää älykkäästi, kumpaa metodia käyttää millekin mittauksille.

Tässä analyysissä käytämme SMOGN algoritmin implementaatiossa [8] muuten oletus argumentteja, paitsi $k = 256$ ja $relthresh = 0.7$. Eli algoritmin ylinäytteistäminen ottaa huomioon mittauksen 256 lähintä naapuria KNN algoritmilla, sekä kasvatta yli-/alinäytteistämis rajaa.

2.6 F1 arvo

F1 arvo on monissa tilastollisissa menetelmissä käytetty arviointimenetelmä. F1 arvo on sisäisen tarkkuuden (engl precision) sekä herkkyden (engl. recall) harmoninen keskiarvo. Binäärisessä luokittelussa kaavat ovat:

$$Precision = \frac{tp}{tp + fp}, \tag{8}$$

jossa tp (true positive) on oikein ennustettujen kertojen määrä, ja fp (false-positive) on väärin positiiviseksi ennustettujen kertojen määrä.

$$Recall = \frac{tp}{tp + fn}, \tag{9}$$

jossa tp (true positive) on oikein ennustettujen kertojen määrä, ja fn (false-negatives) on väärin negatiiviseksi ennustettujen kertojen määrä.

$$F1 = 2 * \frac{precision * recall}{precision + recall} \tag{10}$$

Moniluokkaisissa ongelmissa, f1 arvo lasketaan keskiarvoistamalla erilaisilla metodeilla. Tässä analyysissä f1 arvon keskiarvoistamisessa käytetään painottamatonta keskiarvoa.

Mallin ennustuksille lasketaan siis luokkakohtainen f1 arvo, ja ennustusten lopullinen f1 arvo on yksinkertaisesti luokkien f1 arvojen keskiarvo.

3 AINEISTO JA OHJELMISTOT

3.1 Aineiston kuvaus

Data on kerätty vuonna 2008 viinin laadun mallintamista sekä yksittäisten tekijöiden vaikutusten arviointia varten [1].

Taulukoissa 3 ja 4 nähdään pythonin Pandas kirjaston tuottama kuvaus puna- ja valkoviini dataseteistä. Taulukoista nähdään, että data on epätasapainoista, sillä esimerkiksi arvosana voi olla 3 - 9, mutta 25. ja 75. percentiilin ero on vain 1. Kummassakaan datasetissä ei ole puuttuvia arvoja.

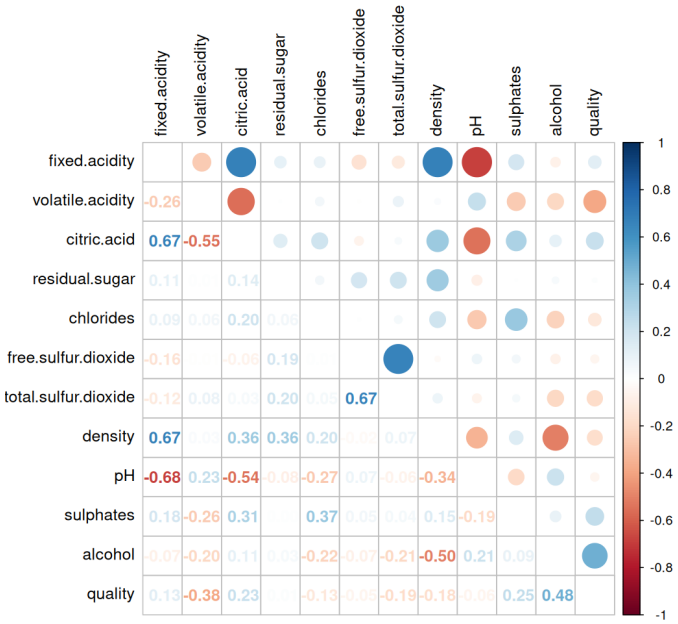
Taulukko 3. Punaviini datan kuvaus

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00	1599.00
mean	8.32	0.53	0.27	2.54	0.09	15.87	46.47	1.00	3.31	0.66	10.42	5.64
std	1.74	0.18	0.19	1.41	0.05	10.46	32.90	0.00	0.15	0.17	1.07	0.81
min	4.60	0.12	0.00	0.90	0.01	1.00	6.00	0.99	2.74	0.33	8.40	3.00
25%	7.10	0.39	0.09	1.90	0.07	7.00	22.00	1.00	3.21	0.55	9.50	5.00
50%	7.90	0.52	0.26	2.20	0.08	14.00	38.00	1.00	3.31	0.62	10.20	6.00
75%	9.20	0.64	0.42	2.60	0.09	21.00	62.00	1.00	3.40	0.73	11.10	6.00
max	15.90	1.58	1.00	15.50	0.61	72.00	289.00	1.00	4.01	2.00	14.90	8.00

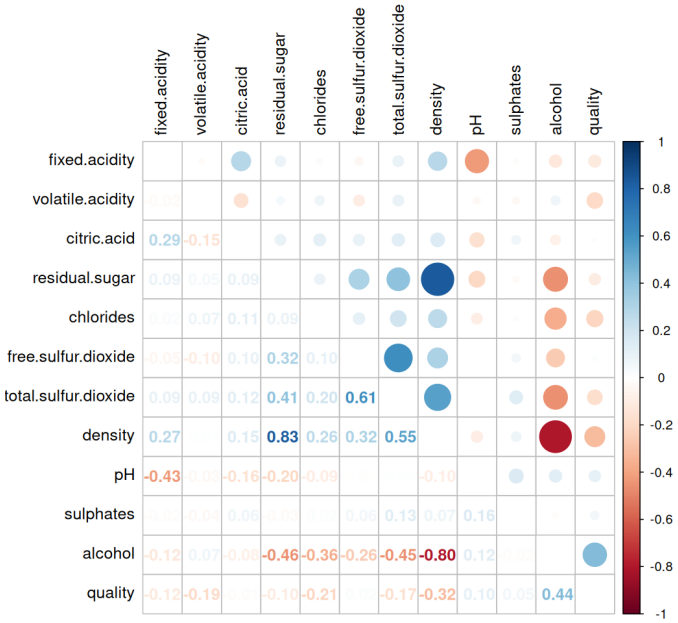
Taulukko 4. Valkoviini datan kuvaus

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00
mean	6.85	0.28	0.33	6.39	0.05	35.31	138.36	0.99	3.19	0.49	10.51	5.88
std	0.84	0.10	0.12	5.07	0.02	17.01	42.50	0.00	0.15	0.11	1.23	0.89
min	3.80	0.08	0.00	0.60	0.01	2.00	9.00	0.99	2.72	0.22	8.00	3.00
25%	6.30	0.21	0.27	1.70	0.04	23.00	108.00	0.99	3.09	0.41	9.50	5.00
50%	6.80	0.26	0.32	5.20	0.04	34.00	134.00	0.99	3.18	0.47	10.40	6.00
75%	7.30	0.32	0.39	9.90	0.05	46.00	167.00	1.00	3.28	0.55	11.40	6.00
max	14.20	1.10	1.66	65.80	0.35	289.00	440.00	1.04	3.82	1.08	14.20	9.00

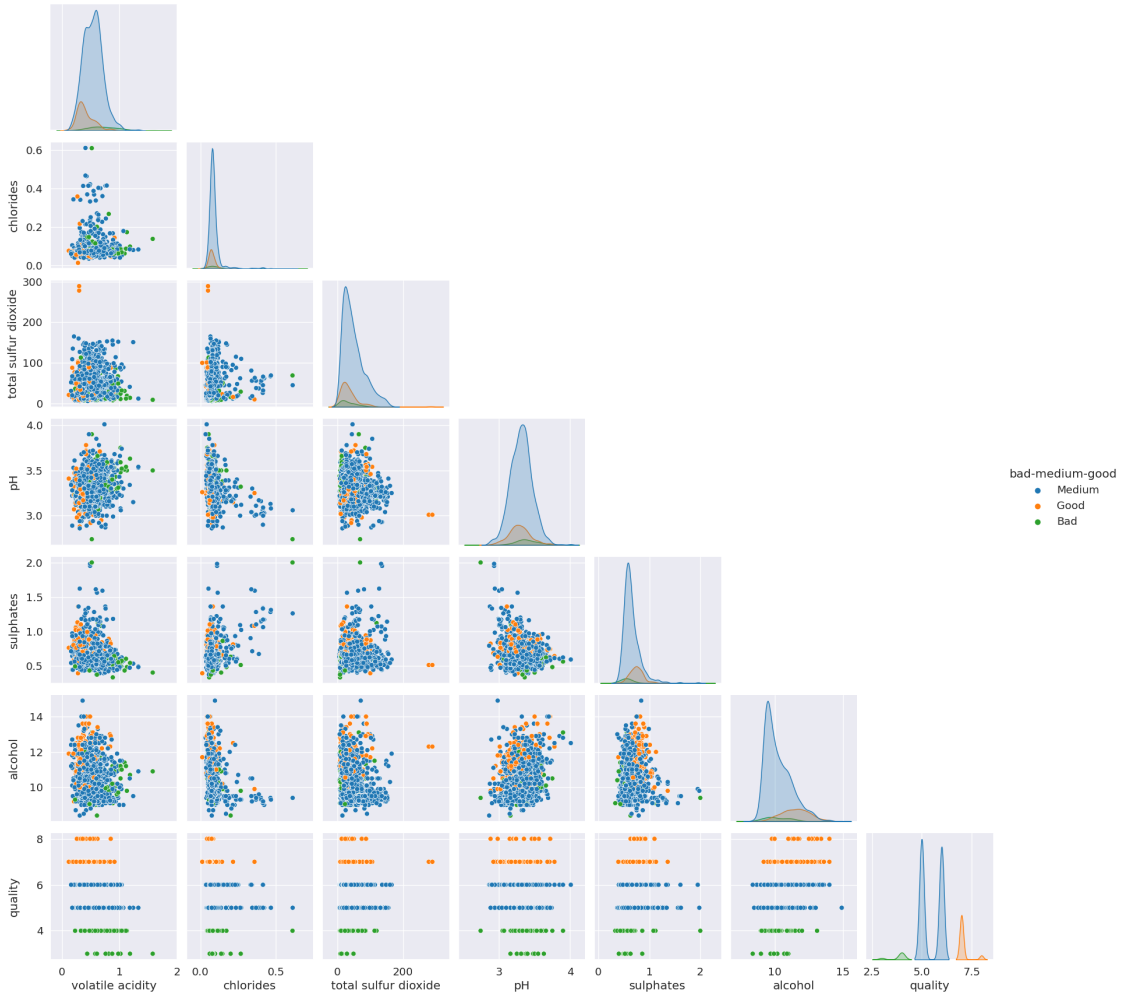
Kuvissa 1 ja 2 on esitetty muuttujien välistä korrelaatiota korrelaatiomatriisein. Kuvasta 2 nähdään, että valkoviinin tiheys korreloi vahvasti kahden muun selittävän muuttujan kanssa. Vahva korrelaatio voi tuottaa ongelmia regressioanalyysissä, joten tiheyden sisällyttämistä malliin vältetään.



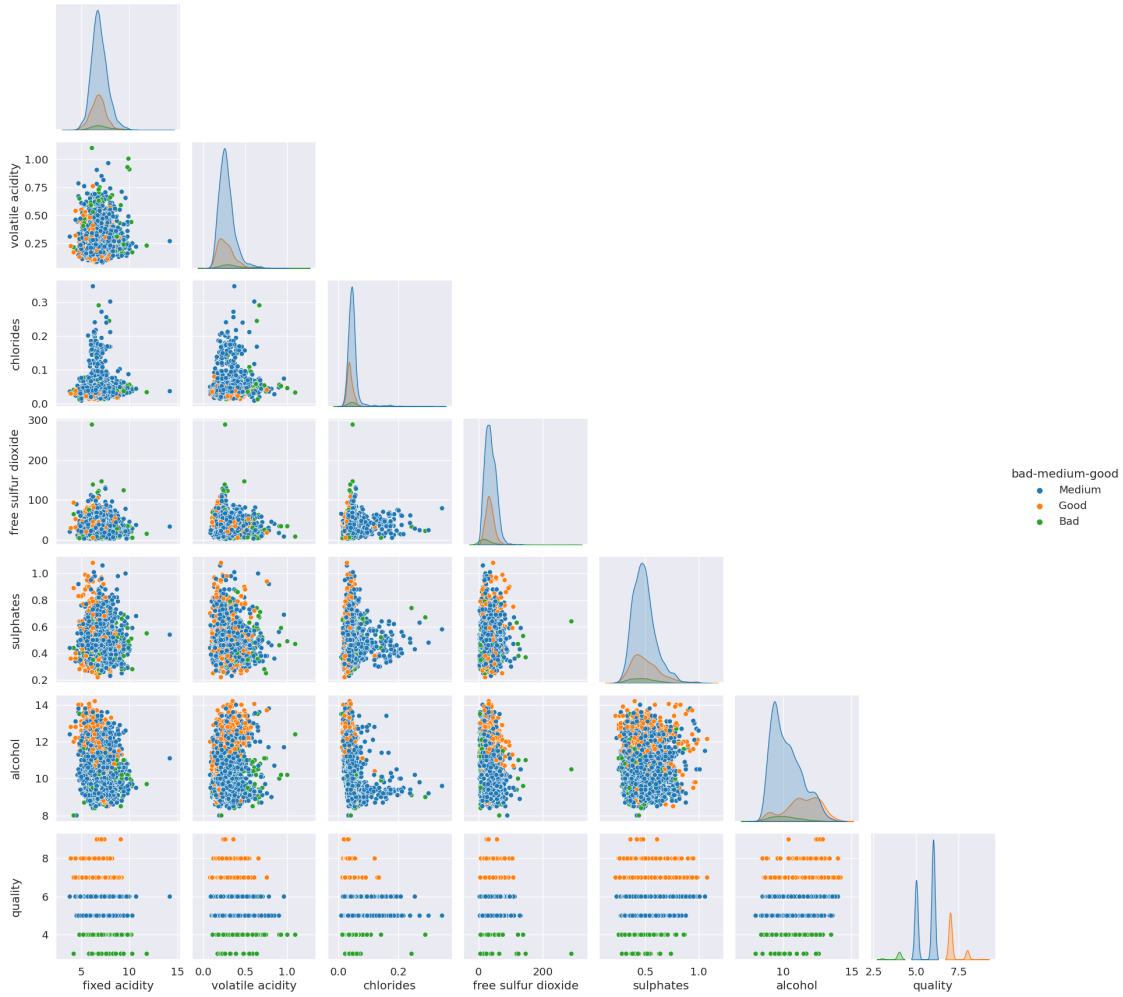
Kuva 1. Korrelaatiomatriisi punaviini



Kuva 2. Korrelaatiomatriisi valkoviini



Kuva 3. Punaviinin muuttujien jakauma. Viinin laadut on luokiteltu huonoihin (0-4, sininen), keskivertoihin (5-6, oranssi), sekä hyviin (7-10, vihreä). Kaikkia muuttujia ei tässä kuvassa näy, sillä muuten kuva olisi liian pieni.



Kuva 4. Valkoviinin muuttujien jakauma. Viinin laadut on luokiteltu huonoihin (0-4, sininen), keskivertoihin (5-6, oranssi), sekä hyviin (7-10, vihreä). Kaikkia muuttujia ei tässä kuvassa näy, sillä muuten kuva olisi liian pieni.

3.2 Ohjelmistot

Python 3.8.10

R 4.2.2

3.3 Aineiston valmistelu

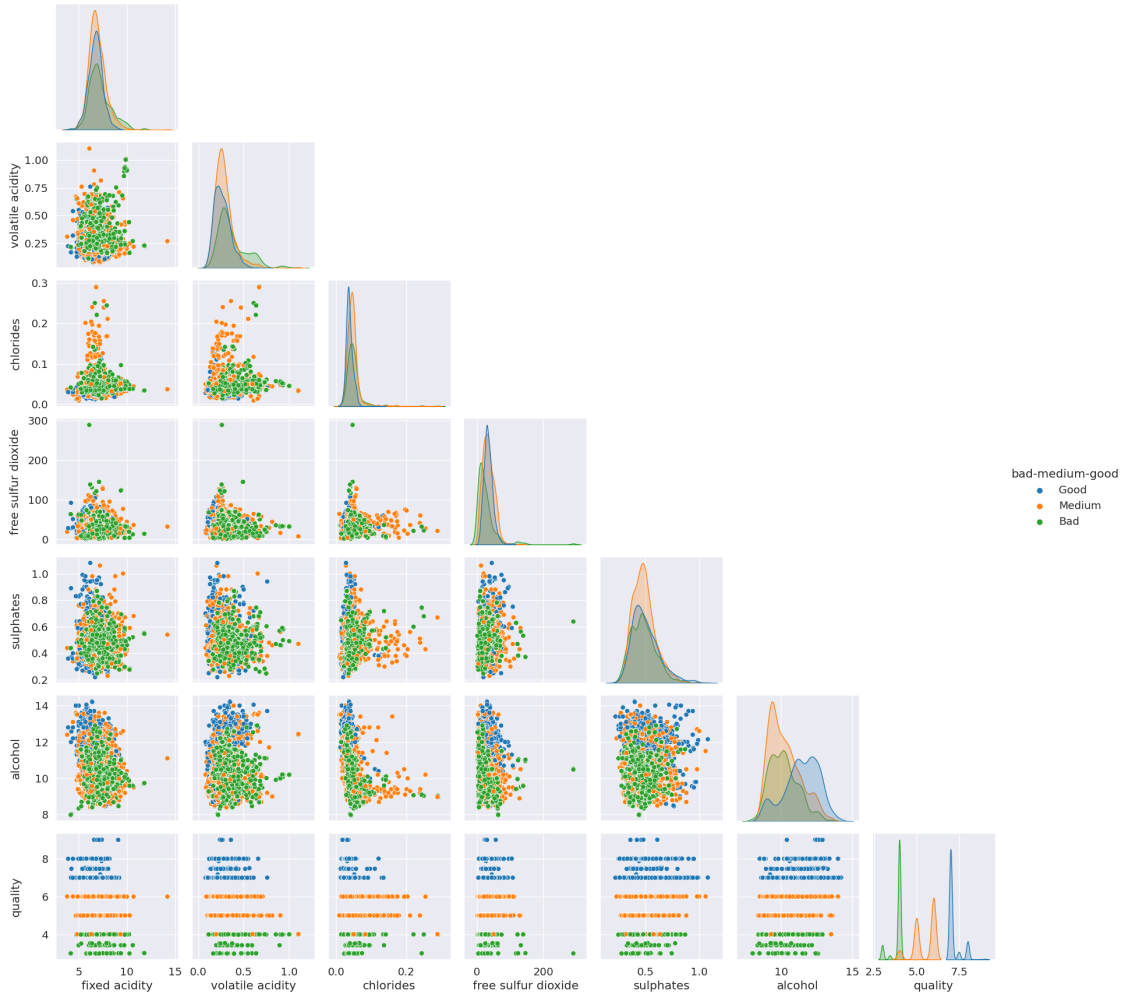
3.3.1 Epätasapainoisuus

Eräs analyysissä ongelmia aiheuttava asia, oli datan epätasapainoisuus. Datassa oli huomattavasti enemmän keskiverto viinejä, kuin hyviä tai huonoja viinejä. Tällöin perinteisillä menetelmillä malli biasoituu ennustamaan yleisiä arvoja, ja sen yleistämiskyky on heikko.

Perinteisesti epätasapainoisessa datassa voidaan joko korjata mallia esim. painottamalla ennustuksen loss funktiota epätasapainoisuuden mukaan ja/tai korjata dataa lisäämällä synteettisiä mittauksia tai yli- tai ali näytteistämällä mittauksia.

Tässä analyysissä yleistämiskykyä mitataan painottamattomalla f1 arvolla. Yleistämiskykyä korjataan muokkamalla dataa, sillä koneoppimismallien muokkaus ei joko tuonut yhtä hyviä tuloksia, tai kirjallisuudesta löytyneille menetelmille ei löytynyt valmista implementaatiota mallien muokkaukselle, kuten *balanced mean squared error* funktiota Tensorflowsta. Dataa muokattiin joko kopioimalla harvinaisten laatujen mittauksia, tai synteettisesti SMOGN algoritmilla [9]. Mitattuun laatuun lisättiin kohinaa satunnaisesti -0.125 ... 0.125, jotta SMOGN algoritmin implementaatio [8] toimi luotettavasti (kuvat 5 ja 6).

Kuva 5. Punaviinin muuttujien jakauma SMOGN käsittelyn jälkeen. Vrt. alkuperäiseen jakaumaan (Kuva 3) käsitellyssä datassa on huomattavasti paremmin jakautuneet ennustettavan muuttujan arvot. Viinin laadut on luokiteltu huonoihin (0-4, vihreä), keskivertoihin (5-6, oranssi), sekä hyviin (7-10, sininen). Kaikkia muuttujia ei tässä kuvassa näy, sillä muuten kuva olisi liian pieni.



Kuva 6. Valkoviinin muuttujien jakauma SMOGN käsittelyn jälkeen. Vrt. alkuperäiseen jakaumaan (Kuva 4) käsitellyssä datassa on huomattavasti paremmin jakautuneet ennustettavan muuttujan arvot. Viinin laadut on luokiteltu huonoihin (0-4, vihreä), keskivertoihin (5-6, oranssi), sekä hyviin (7-10, sininen). Kaikkia muuttujia ei tässä kuvassa näy, sillä muuten kuva olisi liian pieni.

4 TULOKSET

4.1 Lineaarinen regressio

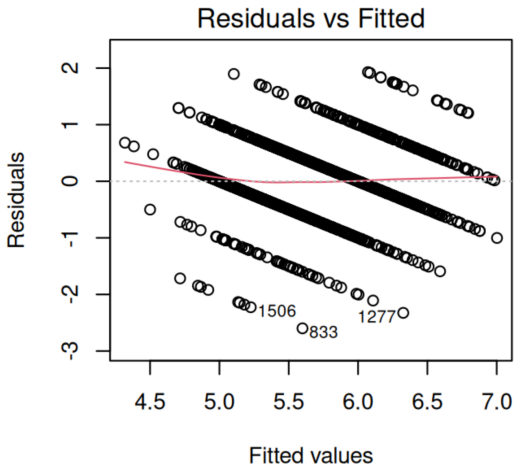
4.1.1 Punaviini

Punaviinille suoritettiin regressioanalyysi käyttäen kuutta(6) selittävää muuttujaa. Regressioanalyysin tulokset on esitetty taulukossa 5.

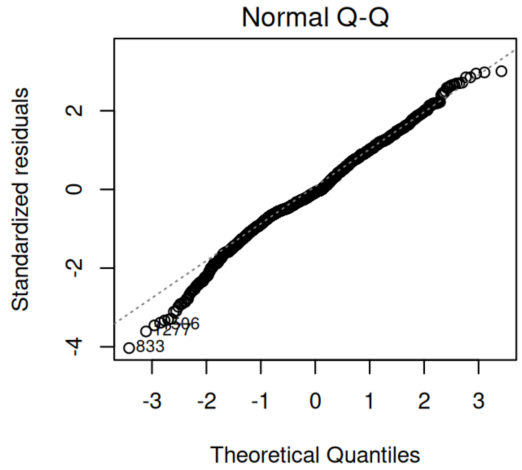
Taulukko 5. Lineaarisen regressioanalyysin tulokset punaviini

	Estimate	Std. Error	t-statistic	p-value
Intercept	4.294	0.398	10.796	0.000
Alcohol	0.300	0.017	17.773	0.000
Volatile Acidity	-1.001	0.101	-9.918	0.000
Sulphates	0.885	0.110	8.080	0.000
Chlorides	-1.965	0.396	-4.960	0.000
pH	-0.469	0.116	-4.055	0.000
Total Sulfur Dioxide	-0.002	0.001	-4.628	0.000
Residual standard error: 0.6455 on 1590 degrees of freedom				
Multiple R-squared: 0.360				
Adjusted R-squared: 0.357				
F-statistic: 149.0 on 6 and 1590 DF				
p-value: 0.000				

Tutkimalla kuvaajia 7 ja 8 voidaan nähdä, että residuaalit eivät ole normaalijakautuneet ja malli sisältää mahdollisesti epälineaarisuutta.



Kuva 7. Residuals vs Fitted kuvaaja punaviini



Kuva 8. Kvantiilikuvio punaviini

4.1.2 Valkoviini

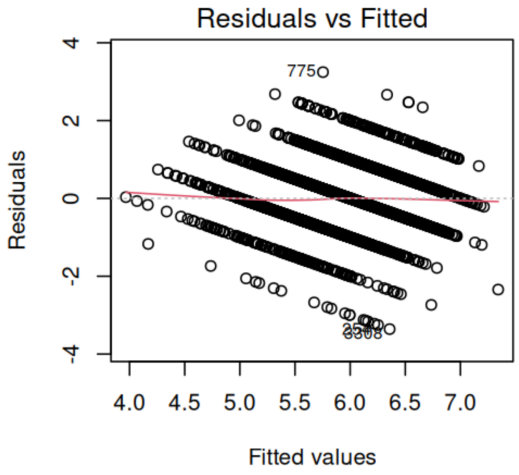
Valkoviinille suoritettiin regressioanalyysi käyttäen kuutta(6) selittävää muuttujaa. Regressioanalyysin tulokset on esitetty taulukossa 6.

Taulukko 6. Lineaarisen regressioanalyysin tulokset valkoviini

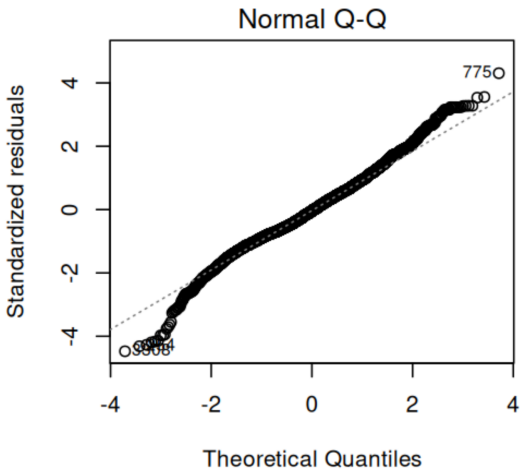
	Estimate	Std. Error	t-statistic	p-value
Intercept	2.407	0.165	14.565	0.000
Fixed acidity	-0.069	0.013	-5.246	0.000
Volatile Acidity	-2.010	0.109	-18.439	0.000
Residual Sugar	0.024	0.003	9.360	0.000
Free Sulfur Dioxide	0.004	0.001	6.251	0.000
Sulphates	0.412	0.095	4.347	0.000
Alcohol	0.380	0.010	37.806	0.000

Residual standard error: 0.7544 on 4885 degrees of freedom
Multiple R-squared: 0.274
Adjusted R-squared: 0.273
F-statistic: 307.2 on 6 and 4885 DF
p-value: 0.000

Tutkimalla kuvaajia 9 ja 10 voidaan nähdä, että residuaalit eivät ole normaalijakautuneet ja malli sisältää mahdollisesti epälineaarisuutta.



Kuva 9. Residuals vs Fitted kuvaaja valkoviini



Kuva 10. Kvantiilikuvio valkoviini

4.2 Logistinen regressio

Viineille toteutettiin logistinen regressioanalyysi ja selittävät muuttujat malleille valittiin korrelaation ja merkitsevyyden perusteella. Molemmista dataseteistä poistettiin nollarivit ja data jaettiin treeni- ja testidataan 80/20 painotuksella. Vaikutus laatuun -arvo kertoo, kuinka paljon todennäköisempää viinille on kuulua astetta parempaan laatuluokkaan selittävän muuttujan kasvaessa yhdellä yksiköllä. Datan epätasapainoisuus näkyi tuloksissa epävarmuutena.

4.2.1 Punaviini

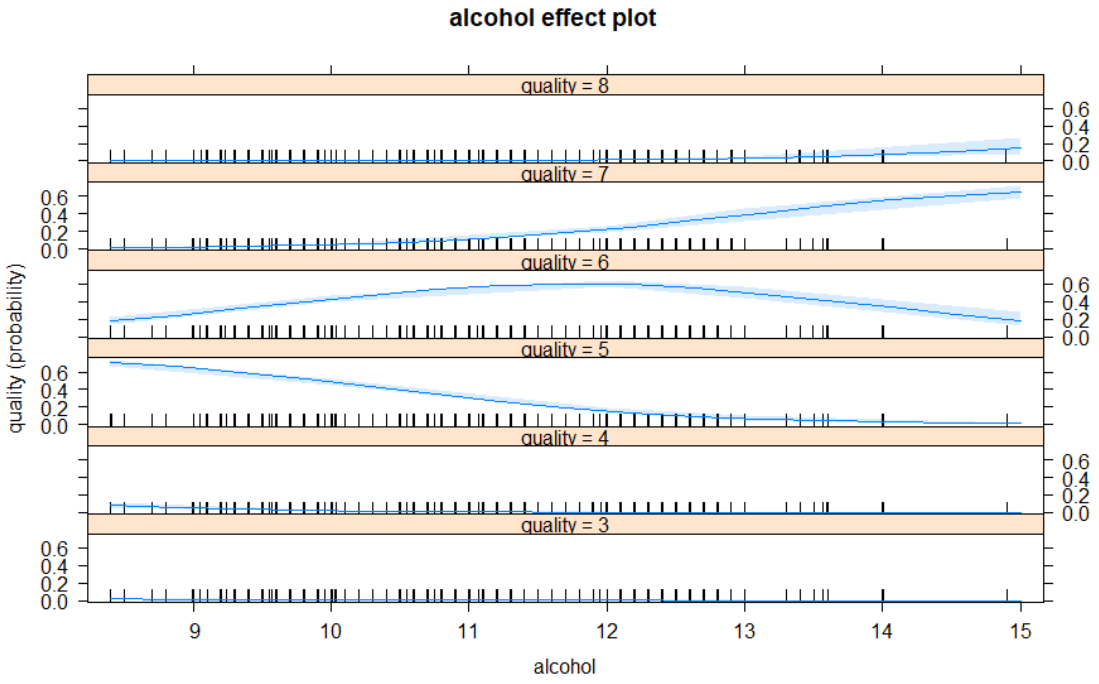
Punaviinille luodun logistisen mallin muuttujien kerroinestimaatit, kyseisen muuttujan vaikutus laatuluokkien suhteeseen, kun muut muuttujat ovat vakioita ja laatuluokkien kynnysarvo löytyvät taulukosta (7). Sulphates-muuttujan vaikutus laatuun on merkittävän suuri verrattuna muihin ja tätä voidaan selittää juuri datan epätasapainoisuudella.

Taulukko 7. Logistisen regression tulokset punaviinille

Muuttuja	Estimaatti	Vaikutus laatuun	Luokka	Kynnysarvo
Volatile acidity	-2.819	0.060	3-4	3.297
Chlorides	-4.566	0.010	4-5	5.243
Total sulfur dioxide	-0.009	0.991	5-6	8.861
Sulphates	3.927	50.762	6-7	11.678
Alcohol	0.8526	2.346	7-8	14.778

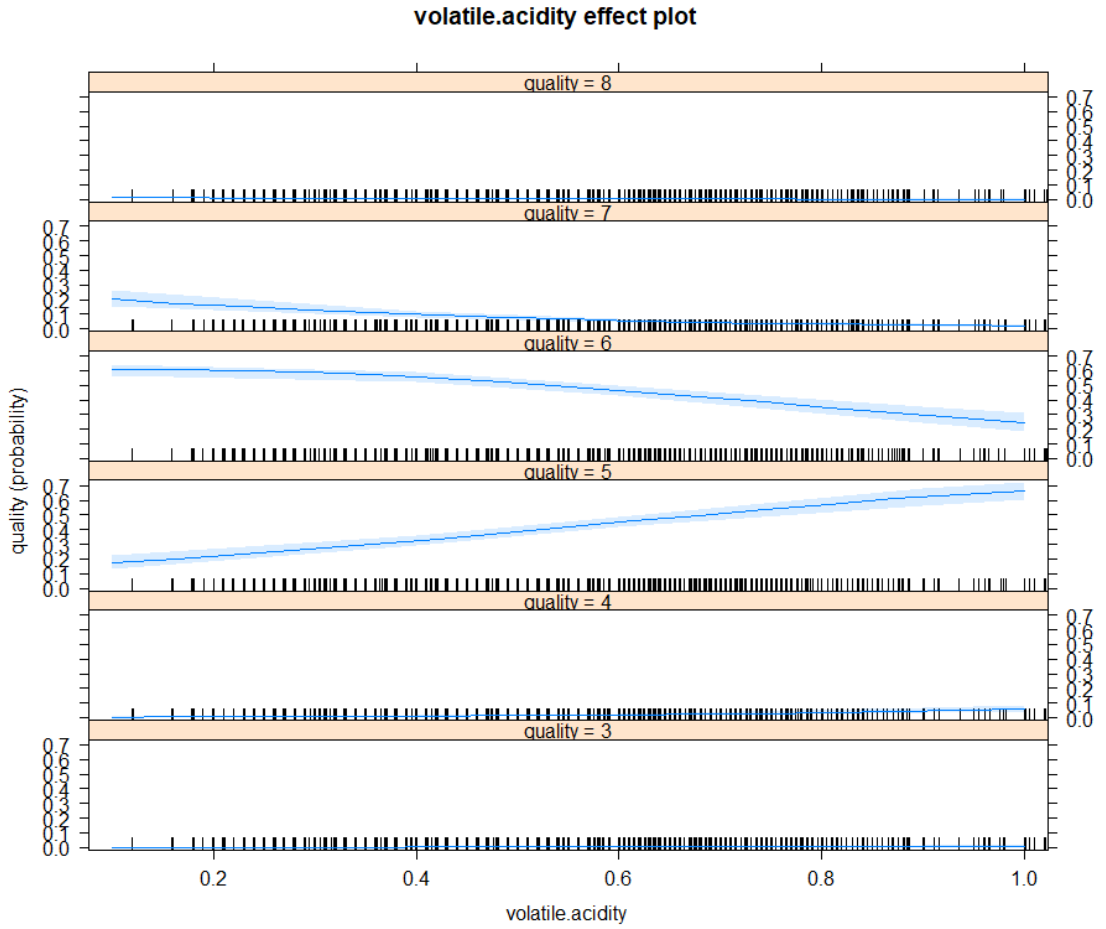
Tarkastellaan seuraavaksi muuttujien alkoholi (engl. alcohol) ja haihtuvat hapot (engl. volatile acidity) vaikutusta punaviiniin laatuun graafisesti.

Logistisen regression tuloksista alkoholin vaikutuskertoimeksi saatin 2.607, mikä käytännössä tarkoittaa että yhden tilavuusprosentin kasvu alkoholissa antaa viinille 2.607 kertaisen mahdollisuuden olla suuremmassa laatuluokassa. Tämä vaikutus nähdään alla olevassa kuvassa todennäköisyyden kasvaessa yhdessä alkoholin kanssa (kuva 11). Datan epätasapainoisuuden vuoksi ääriluokissa ennustaminen on vaikeaa.



Kuva 11. Alkoholin määrän vaikutus punaviinin eri laatuluokissa

Haihtuvien happojen vaikutus taas negatiivinen viinin laatuun arvolla 0.035. Haihtuvilla hapoilla ei siis ole juurikaan positiivista vaikutusta viinin laatuun, enemmänkin negatiivinen (kuva 12).



Kuva 12. Haihtuvien happojen vaikutus punaviinin eri laatuluokissa

4.2.2 Valkoviini

Luotiin samanlainen malli valkoviinille. Tällä kertaa datan ollessa laajempi saatiin tasaisempia tuloksia, esitettu taulukossa (8).

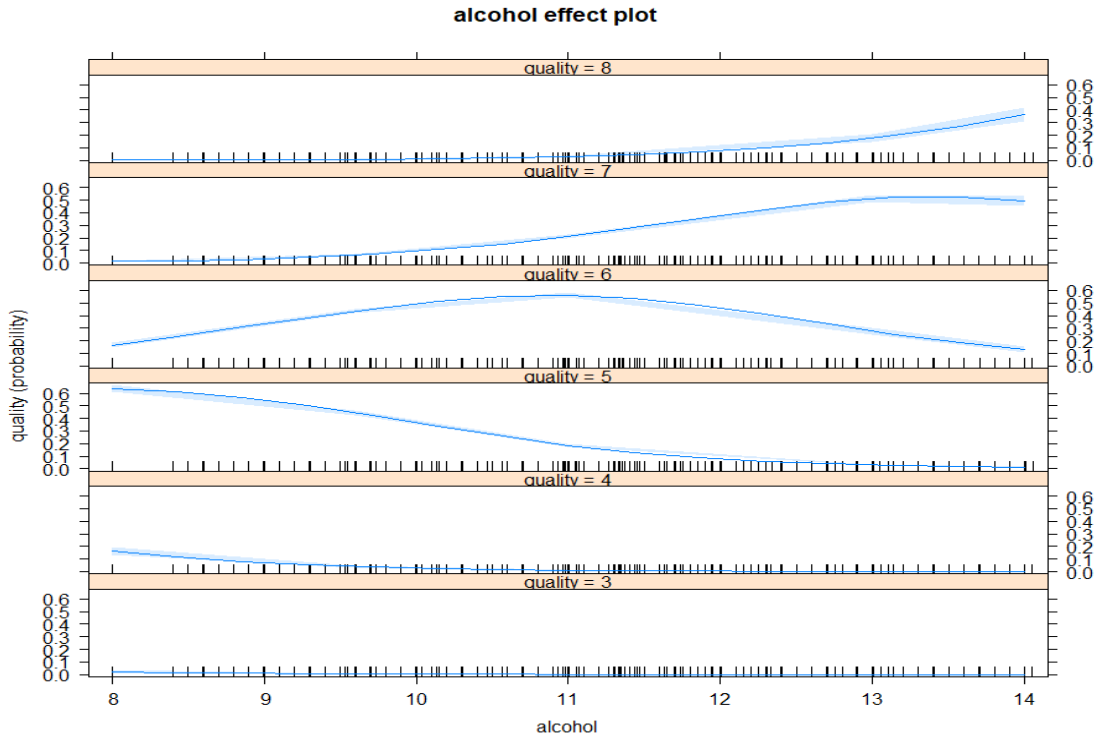
Taulukko 8. Logaritmisen regression kertoimet valkoviinille, kertoimen vaikutuksen suuruusluokka ja laatuluokan kynnysarvot

Muuttuja	Estimaatti	Vaikutus laatuun
Fixed acidity	-0.195	0.823
Volatile acidity	-5.729	0.003
Residual sugar	0.076	1.079
Sulphates	1.248	3.482
Alcohol	0.973	2.645

Luokka	Kynnysarvo
3-4	2.156
4-5	4.480
5-6	7.486
6-7	10.041
7-8	12.373

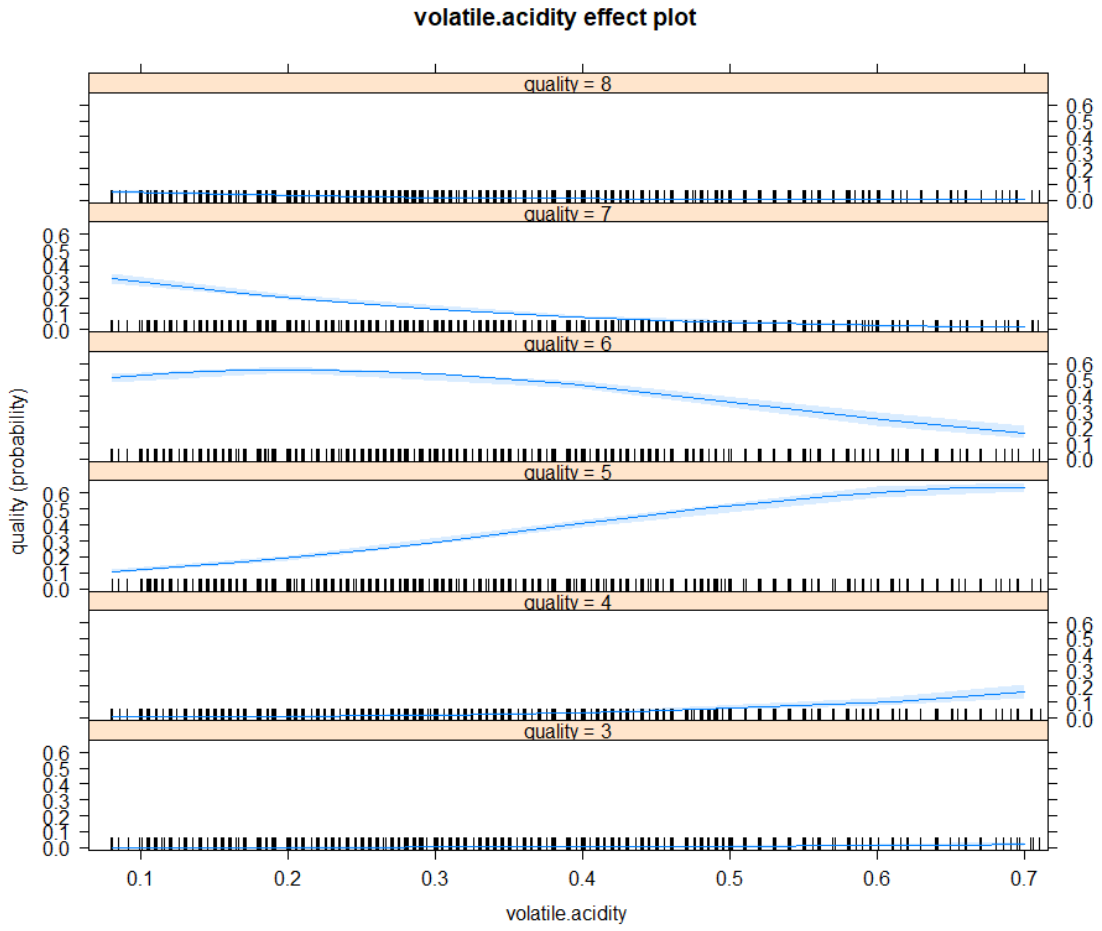
Tarkastellaan myös valkoviinin osalta muuttujien alkoholi (engl. alcohol) ja haihtuvat hapot (engl. volatile acidity) vaikutusta valkoviiniin laatuun graafisesti.

Molemmissa malleissa alkoholin vaikutuskerrorin laatuun on samaa luokkaa. Kuitenkin nähdään (kuva 13) että, kun dataa on enemmän ja se on punaviiniin verrattuna paremmin jakautunut (ei mikään hyvä vieläkään) nähdään parempia tuloksia myös ääriluokissa, joita on määrällisesti vähemmän.



Kuva 13. Alkoholin määrän vaikutus valkoviinin eri laatuluokissa

Haihtuvien happojen vaikutuskertoimen valkoviinin kohdalla ollessa erittäin pieni 0.003, näkyy vaikutus myös laajemmin valkoviinissä (kuva 14).



Kuva 14. Haihtuvien happojen vaikutus valkoviinin eri laatuluokissa

4.3 Laadun ennustus

4.3.1 Yleinen lähestymistapa

Viinin laadun ennustusta kohtelimme regressio ongelmana, vaikka ennustettava muuttuja ei ollut jatkuva, vaan viinin laatu oli pyöristetty lähimpään kokonaislukuun. Ongelmaa olisi myös voinut kohdella luokittelu ongelmana, mutta perinteisesti luokittelussa luokkien välillä ei ole selkeää etäisyyttä ennen mallin opettamista, kun taas regressiossa ennustettavan muuttujan arvoilla on selkeä etäisyys toisistaan. Alustavat tulokset luokittelevista menetelmistä olivat myös heikompia kuin regressiossa.

Laadun ennustuksen hyvyyttä mittasimme kuitenkin luokittelussa käytetyillä menetelmillä, sillä ne todettiin selkeämmiksi. Regressiomallin ennustuksen katsoimme olevan oikea, mikäli ennustus pyöristyi oikeaan kokonaislukuun. Sallittu absoluuttinen virhe oli siis 0.5. Mittareina käytimme tarkkuutta, sekä painottamatonta f1 arvoa. Tulostemme vertailussa aikaisempiin tuloksiin, käytimme myös muita mittareita.

Regressiomalleina käytimme lineaarista mallia, neuroverkkoa, sekä satunnaismetsää. Jaoimme datan testi- ja treenidataan. Testidatan koko oli 25 % koko datasetistä. Erillisellä testidatalla varmistetaan, ettei epälineaarinen malli ylisovita dataan.

4.3.2 Tulokset

Tuloksista (taulukko 9) huomataan, että sekä paras f1 arvo, että paras tarkkuus saavutetaan molemmilla dataseteillä Satunnaismetsällä. Ylinäytteistäminen kasvattaa kaikissa tapauksissa f1 arvoa, mutta suurimmassa osassa tapauksia vain marginaalisesti.

Mitattu tarkkuus saattaa antaa hieman skeptisen kuvan, mutta tässä pitää huomioida se, että malli on regressiomalli, jossa ennustus todetaan oikeaksi JOS ennustettu arvo pyöristyy oikeaan kokonaislukuun, eli absoluuttinen virhe on suurempi kuin 0.5, joka on kuitenkin vielä huomattavan lähellä viinin oikeaa laatua. Mm. MEA:ssa huomataan kaikkien mallien olevan keskimäärin hyvin lähellä oikeaa tulosta.

HUOM. Tulokset joissa on käytetty datan transformointia, saattavat sisältää virhettä, sillä informaatio testi datasta on saattanut epäsuorasti vuotaa treenidataan transformoinnin yhteydessä.

Taulukko 9. Koneoppimismallien tulokset, normalisaatiota (5) ei ole lueteltu erikseen 'Transforamtion' kohtaan, vaan se on tehty kaikille muuttujille, kaikissa tapauksissa.

Wine	Model type	Oversampling	Transformation	Mean Absolute Error	Accuracy (%)	F1 score (macro)
Red	Linear	NaN	Boxcox	-0.0300	62.00	0.275
Red	Linear	Smogn	Standard	0.2200	53.50	0.311
Red	Neural Net	NaN	Standard	-0.0600	62.75	0.312
Red	Neural Net	Manual rares	Standard	0.0550	60.75	0.370
Red	Random Forest Regression	NaN	Boxcox	0.0098	69.20	0.330
Red	Random Forest Regression	Manual rares	NaN	0.0300	67.00	0.351
White	Linear	NaN	Standard	-0.0200	51.10	0.230
White	Linear	Smogn	Standard	-0.1800	44.10	0.280
White	Neural Net	NaN	Standard	-0.0500	56.90	0.280
White	Neural Net	Smogn	Standard	-0.0500	54.80	0.280
White	Random Forest Regression	NaN	NaN	0.0100	68.80	0.430
White	Random Forest Regression	Smogn	NaN	0.0700	64.10	0.480

5 JOHTOPÄÄTÖKSET

Kappale on jaettu kahteen osaan. Ensin analysoidaan tulokset. Tämän jälkeen käydään läpi ehdotuksia kuinka malleja ja menetelmiä voitaisiin parantaa tulevaisuuden tutkimuksissa.

5.1 Tulosten analysointi

Lineaarisesta ja logistisesta regressiomalleista ei saatu kovinkaan hyviä ennusteita yksittäisen viinin laadun luokitteluun. Mallien tulokset kuitenkin täydentävät toisiaan ja antavat yhdessä tutkittuna hyvän osviitan siitä, mitkä fysikokemialliset muuttujat vaikuttavat viinin laatuun ja millä suuruudella.

Punaviinille (taulukot 5 ja 7) ja valkoviinille (taulukot 6 ja 8) saatuja tuloksia tarkkailtaessa huomataan, että viinien laatua selittävät muuttujat ovat samoja ja niiden kerroinestimaatit samaa suuruusluokkaa samalla etumerkillä. Täytyy kuitenkin muistaa, että tuloksia ei voida suoraan verrata keskenään mallien erilaisuuksien takia.

Laadun ennustuksessa eri koneoppimismalleilla päästään hyvään tulokseen monella mittarilla (taulukko 9). Satunnaismetsillä saadut tarkkuudet ylittävät alkuperäisessä tutkimuksessa SVM:llä saadut tarkkuudet.

5.2 Jatkotutkimus

SMOGN algoritmille olisi varmasti löytynyt jokaiselle käytölle kustomoidut ja paremmat parametrit. Nyt SMOGN algoritmia käyttäessä ali-/ylinäytteistämisen suhde näyttää silmämääräisesti liian suurelta varsinkin valkoviinin datasetissä (Kuva 6).

Neuroverkkojen arkkitehtuuri, sekä satunnaismetsän parametrit pitäisi optimoida jokaista eri testiä kohden. Neuroverkkojen tarkkuus vastaa alkuperäisessä tutkimuksessa saatuja tuloksia.

Regressionmalleissa todettiin olevan epälineaarisuutta. Sopivin muuttuja muutoksin epälineaarisuudesta voitaisiin päästä eroon ja mallin selitysastetta parantaa.

LÄHTEET

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *In Decision Support Systems, Elsevier, 47(4):547-553*, 2009.

[2] Future Market Insights. Wine market outlook - 2022-2032, Aug 2022. Accessed on 2022-19-12.

[3] G. Schamel. Individual and collective repuatation indicators of wine quality. *SSRN Electronic Journal*, 2000.

[4] E. Oczkowski. The effect of weather on wine quality and prices: An australian spatial analysis. *Journal of wine economics*, 2016.

[5] A. Agresti and C. Tarantola. Simple ways to interpret effects in modeling ordinal categorical data. 2018. Luettu 2022-12-08.

[6] J. Brownlee. How to avoid data leakage when performing data preparation. https://machinelearningmastery.com/data-preparation-without-data-leakage/, Aug 2020. Accessed on 2022-19-12.

[7] L. Torgo, R. Ribeiro, P. Branco, and P. Pfahringer. Smote for regression. 2013.

[8] Nicholas Kunz. SMOGN: Synthetic minority over-sampling technique for regression with gaussian noise, 2020.

[9] Luis Torgo, Bartosz Krawczyk, Paula Branco, and Nuno Moniz. Smogn: a pre-processing approach for imbalanced regression. 2017.