**SVEUČILIŠTE JOSIPA JURJA STROSSMAYERA U OSIJEKU**

**FAKULTET ELEKTROTEHNIKE, RAČUNARSTVA I INFORMACIJSKIH TEHNOLOGIJA OSIJEK**

**Diplomski studij**

**PREDVIĐANJE KVALITETE ČOKOLADE**

**SEMINARSKI RAD**

**RAČUNARSTVO USLUGA I ANALIZA PODATAKA**

**Željko Sabo**

**Osijek, 2019.**

**SADRŽAJ**

[1. UVOD 1](#_Toc17118639)

[2. OPIS 2](#_Toc17118640)

[3. KORIŠTENI PODACI 3](#_Toc17118641)

[4. KORIŠTENI POSTUPCI STROJNOG UČENJA 4](#_Toc17118642)

[5. MODEL STROJNOG UČENJA 5](#_Toc17118643)

[5.1. Planiranje i izrada modela 5](#_Toc17118644)

[5.2. Testiranje i evaluacija dobivenih rezultata 6](#_Toc17118645)

[5.3. Izrada web servisa i korištenje API-a 10](#_Toc17118646)

[5.4. Testiranje izmijenjenog skupa podataka 13](#_Toc17118647)

[6. OPIS PROGRAMSKOG RJEŠENJA 14](#_Toc17118648)

[6.1. Forma za unos parametara i prikaz rezultata 14](#_Toc17118649)

[7. ZAKLJUČAK 16](#_Toc17118650)

[LITERATURA 17](#_Toc17118651)

[POVEZNICE 17](#_Toc17118652)

# UVOD

U ovom projektu obrađuje se problem predviđanja kvalitete čokolade pomoću modela strojnog učenja. Aplikacija treba omogućiti unos parametara na temelju kojih se čokolada, ovisno o parametrima, klasificira u jednu od tri kategorije a to su loša, dobra i izvrsna čokolada.

Pošto je čokoladu potrebno svrstavati u klase ovisno o njenoj kvaliteti, u ovom projektu je korištena metoda klasifikacije pomoću koje se čokolada ovisno o svojoj kvaliteti svrstava određenu klasu.

U seminaru će biti opisani korišteni skup podataka, postupci strojnog učenja, izrada modela, objavljivanje web servisa te opis programskog rješenja.

# OPIS

U cijelome svijetu čokolada je najpopularnija slastica koja se konzumira. Najčešće se konzumira u obliku čokoladnih pločica kao „obična“ čokolada, no često se koristi i u kuhinji, prilikom spremanja različitih poslastica kao što su na primjer kolači.

Naravno, postoje različiti brendovi čokolade, te sukladno tome, nisu svi brendovi jednako dobri. Čokolade mogu imati različite postotke kakaa u sebi, mogu isto tako biti i mliječne te samim tim u sebi sadržavati određenu količinu mlijeka, no isto tako mogu i da sadrže mnoge druge dodatke poput karamela, kikirikija, keksa i brojnih drugih. Također ne koriste svi proizvođači ista zrna kakaa i nisu svaka zrna jednako dobra, te se ovaj projekt bazira na prepoznavanju i klasifikaciji čokolade na osnovu tih, i brojnih drugih parametara.



**Slika 2.1. –** Prikaz različitih brendova čokoladnih pločica

O ovom projektu biti će korišteni parametri koje korisnik sam unosi, te pomoću kojih se rezultat može klasificirati u (kao što je prethodno spomenuto) jednu od tri klase (razreda) a to su loše čokolade, dobre čokolade i izvrsne čokolade. Parametri koji će se koristiti za klasifikaciju su:

* Naziv tvrtke
* Postotak kakaa (vrijednost od 0 do 1)
* Vrsta zrna

# KORIŠTENI PODACI

Kako bi aplikacija uspješno obavljala svoje zadatke, potrebo je izraditi model strojnog učenja koji će na temelju podataka koji su mu predani moći klasificirati rezultat u jedan od tri razreda koji su modelu dostupni za klasifikaciju (razred loših, dobrih i izvrsnih čokolada).

Najvažniji dio prilikom izrade modela strojnog učenja je bio pronalazak i pred obrada podataka na temelju kojih će se model trenirati. Sve parametre koji su na bilo koji način mogli negativno utjecati na ishod ili koji su bili nepotrebni su izbačeni kako bi se dobio model koji bi se mogao koristiti u što realnijim uvjetima.

Podatkovni skup koji je korišten je preuzet sa internetske stranice Kaggle, te ga je moguće pronaći na linku koji se nalazi u poveznici [1]. Skup podataka se sastoji od 1795 instanci od kojih su izbačene sve gdje je falio jedan od prethodno navedenih parametara kako bi trenirani model radio što točnije. Korišteni set podataka je objavljen 2017. godine od strane Rachael Tatman koja je znanstvenik za podatke u Kaggle-u.



**Slika 3.1. –** Izgled jedne instance

U dobiven podacima osim navedenih parametara koji će se koristiti za klasifikaciju, postoji veliki broj parametara koji se nisu koristili, a to su redi broj instance, mjesto podrijetla zrna, REF (godina prve evaluacije čokolade), godina zadnje evaluacije, ocjena, država u koju je smještena kompanija i država podrijetla zrna.

Redni broj instance je izbačen pošto je u potpunosti nepotreban te na nikakav način ne bi imao utjecaja pri treniranju modela. Bez obzira koristio se redni broj instance ili ne, trenirani model bi imao istu točnost, odnosno pogrešku.

Sljedeći podatak koji je izbačen je mjesto podrijetla zrna čokolade i država podrijetla zrna, koji su izbačeni pošto je taj podatak često nedostupan, te bi korisnici najčešće imali poteškoća s pronalaženjem tih informacija. S obzirom da proizvođači čokolade takve informacije ne stavljaju kao dostupne na svoje proizvode, ukoliko bi željeli evaluirati čokoladu bili bi prisiljeni taj podatak pronaći samostalno putem interneta. Samim tim što bi podatak morali potražiti na internetu bi u evaluaciju unosilo određenu količinu moguće pogreške jer na ni jedan način ne bi mogli znati je li podatak koji smo pronašli točan ili nije.

Zatim REF odnosno godina prve evaluacije čokolade i godina zadnje evaluacije čokolade su izbačene jer postoji veliki broj raznih čokolada koje nikad nisu bile evaluirane, te samim tim ne posjeduju ni REF (godinu prve evaluacije) niti godinu zadnje evaluacije. Veliki broj novih brendova čokolada nastaje svake godine, te se one često ne uspiju zadržati na tržištu. Ukoliko bi htjeli evaluirati neku od novih čokolada koje nisu imale prvu evaluaciju, bili bi prisiljeni unijeti podatke koje nemamo dostupne, te koje ne možemo unijeti.

Ista situacija je i s ocjenom. Ukoliko čokolada nikada nije imala prvu evaluaciju, nikad nije ni mogla dobiti ocjenu, te samim tim bi taj podatak bio nedostupan, ne bi se mogao unijeti, te se ta čokolada ne bi mogla realno evaluirati.

Svi podaci koji su izbačeni su kao i što je i vidljivo iz prethodno objašnjenog su izbačeni radi svoje nedostupnosti i neimanja utjecaja na evaluaciju.

# KORIŠTENI POSTUPCI STROJNOG UČENJA

Nakon što je pronađen odgovarajući skup podataka, te je odrađena obrada korištenog skupa podataka, kao sljedeći korak, bilo je potrebno odabrati postupak strojnog učenja koji će biti najpogodniji, odnosno davati najveću točnost i najmanju pogrešku prilikom evaluacije. Pošto je u ovom projektu bilo potrebno odrediti jeli čokolada loša, dobra ili izvrsna, korištena je metoda klasifikacije, a kao postupak klasifikacije metoda *Multiclass Decision Jungle* pošto su s njom postignuti najbolji rezultati prilikom klasifikacije.

Klasifikacija je postupak u kojem se elementi populacije (u ovom slučaju vrste čokolada) dodjeljuju nekom od razreda u ovisnosti o tome koje kriterije za koji razred elementi koji se razvrstavaju zadovoljavaju. Možda bi čak bilo jednostavnije reći da je klasifikacija metodu učenja pomoću koje se određuje kategorija, tip ili klasa jednog ili više redova podataka.

Najčešća podjela klasifikacija je na to vrše li se klasifikacije na dvije ili više klasa. Ukoliko se klasifikacija vrši na dvije klase tada se radi o binarnoj klasifikaciji. Jednostavan primjer binarne klasifikacije bi bio ukoliko bi se klasifikacija kvalitete čokolade provodila tako da se određuje samo je li čokolada dobra ili loša, no ako je riječ o više od dvije klase kao što je u ovom projektu, onda se radi o više klasnoj klasifikaciji.

## Multiclass Decision Forest

*Multiclass Decision Forest* je skupina metoda učenja namijenjena za klasifikaciju na više klasa, gdje se skupine metoda ne temelje na jednom modelu, već na više njih kako bi dobili bolje rezultate i općenitiji model. Ova metoda funkcionira tako da se izgradi više stabala odlučivanja, te se onda pomoću glasanja određuje izlazna klasa. [2]

## Multiclass Decision Jungle

*Multiclass Decision Jungle* je algoritam koji vraća neobučeni klasifikator, te se model zatim osposobljava korištenjem modela treniranja ili *Tune Model Hyperparameters*. Obučeni model se zatim može koristiti za klasificiranje podataka. Džungle odlučivanja su nedavno proširenje šuma odlučivanja, te se sastoje od skupa acikličkih grafova usmjerenih na odluku. [3]

## Multiclass Neural Network

Neuronska mreža je skup međusobno povezanih slojeve, gdje ulazi predstavljaju prvi sloj koji je povezani sa izlaznim slojem putem acikličkog grafa koji se sastoji od težinskih rubova i čvorova. Između ulaznog i izlaznog sloja može postojati više skrivenih slojeva, a većina zadataka se može veoma lako obaviti uz pomoć jednog ili nekoliko skrivenih slojeva. Odnos između ulaza i izlaza saznaje se iz treninga neuronske mreže na ulaznim podacima. Smjer grafa odvija se od ulaza preko skrivenog sloja i do izlaznog sloja. Svi čvorovi u sloju su težinski rubovi povezani s čvorovima u sljedećem sloju. Da bi se izračunao izlaz mreže za određeni ulaz, izračunava se vrijednost na svakom čvoru u skrivenim slojevima i u izlaznom sloju. Vrijednost se postavlja izračunavanjem težinske sume vrijednosti čvorova iz prethodnog sloja. Zatim se na taj težinski zbroj primjenjuje funkcija aktivacije. [4]

## Multiclass Logistic Regression

*Multiclass Logistic Regression* je klasifikacijska metoda koja generalizira logističku regresiju na probleme više klase, odnosno na probleme sa više od dva moguća diskretna ishoda. To je model koji se koristi za predviđanje vjerojatnosti različitih mogućih ishoda kategorički raspodijeljene zavisne varijable, s obzirom na skup neovisnih varijabli. [5]

# MODEL STROJNOG UČENJA

## Planiranje i izrada modela

Na Azure web platformi napravljen je novi eksperiment u kojem je izrađen model strojnog učenja. Model će na temelju podataka i odabrane klasifikacije predviđati rezultate i njihovu vjerojatnost. Za metodu klasifikacije odabrana je *Multiclass Decision Jungle* metoda.

Kao prvi korak u kreiranju modela, potrebno je dodati skup podataka na osnovu kojeg će model biti treniran. U ovom slučaju skup podataka naziva se „flavors\_of\_cacao.csv“, te je zapisana u CSV (engl. *Comma-Separated Values*) formatu. Korišteni podaci imaju 158 unikatnih kompanija, 30 unikatnih udjela kakaa i 21 unikatnu vrstu zrna.

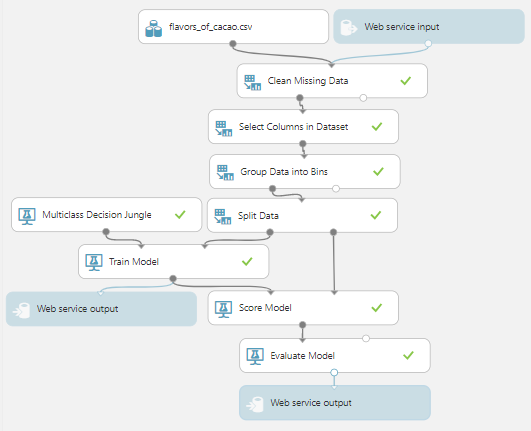
Nakon dodavanja skupa podataka sljedeće što je bilo potrebno napraviti je očistiti i urediti podatke. Uklonjeni su svi redci koji nisu imali sve stupce popunjene. To je napravljeno kako bi se dobio model koji će što točnije vršiti klasifikaciju.

Dalje, podaci su normalizirani i podijeljeni na podatke koji će se koristiti prilikom treniranja modela i na podatke koji će se koristiti za evaluaciju treniranog modela. 70% podataka je korišteno za treniranje modela, dok je 30% korišteno za njegovu evaluaciju.

Prilikom izrade modela korištene su i ispod navedene klasifikacije:

* *Multiclass Decision Forest* [3]
* *Multiclass Neural Network* [4]
* *Multiclass Logistic Regression* [5]

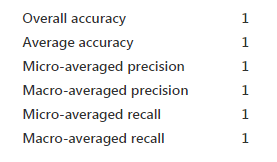
Konačni izgled modela moguće je vidjeti na slici 5.1.



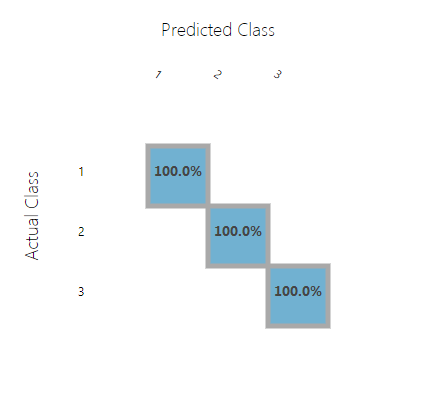
**Slika 5.1. –** Konačni izgled modela

## Testiranje i evaluacija dobivenih rezultata

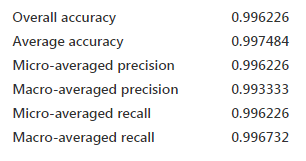
Nakon pokretanja azure servis je obavio testiranje i dobiveni su rezultati. Svi algoritmi klasifikacije koji su korišteni u ovom testu su dali vrlo visoke postotke preciznosti i točnosti, no najtočnija metoda je ipak bila *Multiclass Decision Jungle*. Rezultati *Multiclass Decision Jungle-a* prikazani su na slici 5.2. dok su rezultati ostalih klasifikacija prikazani na slikama 5.3.-5.9.



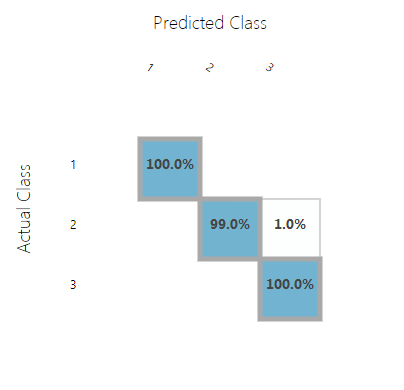
**Slika 5.2. –** Rezultat za *Multiclass Decision Jungle*



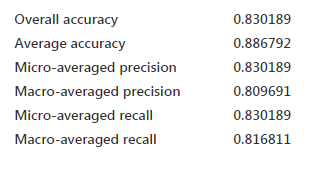
**Slika 5.3. –** Matrica konfuzije za *Multiclass Decision Jungle-a*



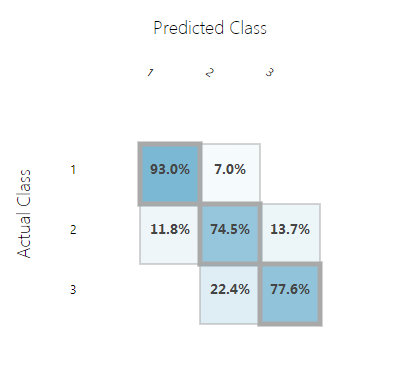
**Slika 5.4. –** Rezultat za *Multiclass Decision Forest*



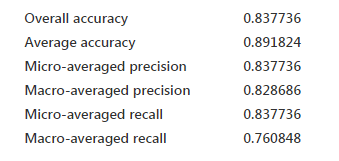
**Slika 5.5. –** Matrica konfuzije za *Multiclass Decision Forest-a*



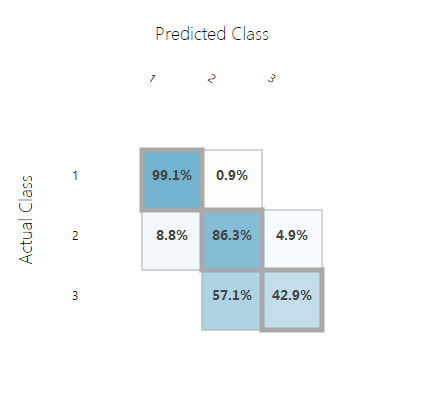
**Slika 5.6. –** Rezultat za *Multiclass Neural Network-a*



**Slika 5.7. –** Matrica konfuzije za *Multiclass Neural Network-a*



**Slika 5.8. –** Rezultati za *Multiclass Logistic Regression*

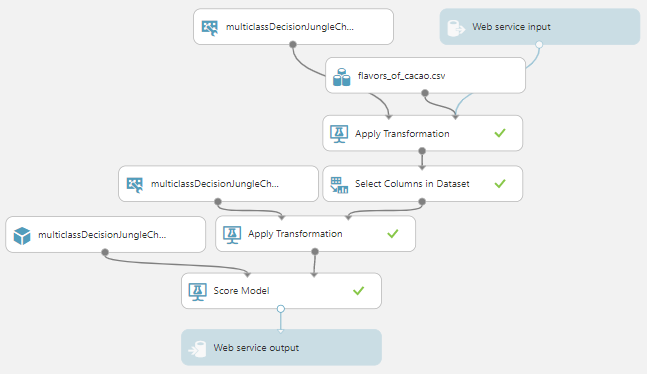


**Slika 5.9. –** Matrica konfuzije za *Multiclass Logistic Regression*

Pomoću dobivenih matrica konfuzije je moguće vidjeti da svi algoritmi imaju veoma visoku točnost klasifikacije, no kao najtočniji se ističe *Multiclass Decision Jungle.* Kod *Multiclass Decision Jungle-a* je moguće vidjeti da su svi podatci točno klasificirani, dok kod ostalih vrsti klasifikacije postoji pogreška. Najveću pogrešku imaju *Multiclass Neural Network* i *Multiclass Logistic Regression*, te ona iznosi blizu 20%. *Multiclass Decision Forest* bi bio isto tako dobar izbor s obzirom da mu pogreška iznosi 1% i to samo prilikom klasifikacije čokolade dobre klase.

## Izrada web servisa i korištenje API-a

Nakon što je završeno sa treniranjem modela sljedeći korak je izrada web servisa koji će preko API-a primati parametre, klasificirati rezultat i vratiti ga kao odgovor. Za izradu web servisa potrebno je odabrati opciju „*Train web service“* pomoću koje se stvara prediktivni eksperiment. (slika 5.10.)



**Slika 5.10. –** Izgled prediktivnog eksperimenta

Nakon što je izrađen prediktivni eksperiment, potrebno ga je pokrenuti i zatim odabrati opciju *„Deploy web service“* nakon čega dobivamo *Request URI* i API ključ. Web servisu pristupamo pomoću dobivenog *Request URI* te API ključa. Uz to je potreban odgovarajuća JSON forma za zahtjev koja se generira iz C# rječnika na koji se dobiva JSON odgovor koji je moguće vidjeti na slici (5.2). JSON je transportni format podatka koji omogućuje jednostavno slanje i primanje podataka putem API-a.

**Programski kôd 5.1. –** Izgled C# rječnika pomoću kojeg se generira JSON zahtjev

{

Inputs = {

System.Collections.Generic.Dictionary<string, System.Collections.Generic.List<System.Collections.Generic.Dictionary<string, string>>>},

GlobalParameters = {System.Collections.Generic.Dictionary<string, string>

}

}

**Programski kôd 5.2. –** Primjer JSON odgovora

{

"Results": {

"output1": [

{

"Company":"Guittard",

"Cocoa Percent":"0.7",

"Bean Type":"Bland",

"Cocoa Percent\_quantized":"1",

"Scored Probabilities for Class "1"":"0.0419902912621359",

"Scored Probabilities for Class "2"":"0.193446601941748",

"Scored Probabilities for Class "3"":"0.764563106796117",

"Scored Labels":"3"

}

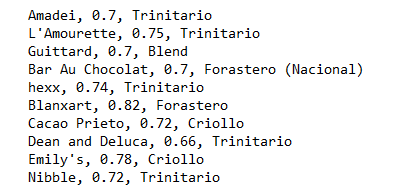
]

}

}

## Testiranje izmijenjenog skupa podataka

U svrhu testiranja aplikacije iz skupa podataka izdvojeno je 10 podataka koji će služiti za potrebe testiranja. Podatci su uneseni u aplikaciju te se je vršila evaluacija aplikacije. Podatci su odabrani nasumično te se mogu vidjeti na slici 5.3.



**Slika 5.3. –** Instance nasumično odabrane za testiranje

Nakon što su sve pripreme obavljene, ručno su testirane izbačene instance. Rezultati su prikazani u matrici konfuzije koja je vidljiva u tablici 5.1.

|  |  |  |  |
| --- | --- | --- | --- |
| n = 10 | Predviđeni:  Loša | Predviđeni:  Dobra | Predviđeni: Izvrsna |
| Stvarni:  Loša | 4 | 0 | 0 |
| Stvarni:  Dobra | 0 | 2 | 0 |
| Stvarni: Izvrsna | 0 | 0 | 4 |

**Tablica 5.1. –** Matrica konfuzije

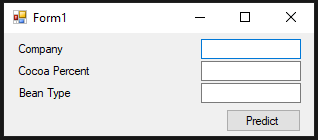
Iz matrice konfuzije je vidljivo da je od 10 čokolada 4 klasificirano kao izvrsne, 2 kao dobre i 4 kao loše čokolade. Iz testiranih instanci je moguće vidjeti da su sve imale različita imena tvrtki koje ih proizvode i različite postotke kaka, no neke su imale istu vrst zrna.

# OPIS PROGRAMSKOG RJEŠENJA

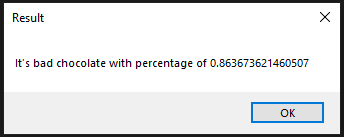
Nakon izrade modela strojnog učenja i objave web servisa potrebno je izraditi aplikaciju za testiranje. Aplikacija treba omogućiti unos parametara koji će biti testirani, te ispis rezultata u obliku određene klase i vjerojatnosti. Aplikacija je izrađena kao *Windows Form* aplikacija u *C#* programskom jeziku, unutar *Visual Studio* programskog okruženja.

## Forma za unos parametara i prikaz rezultata

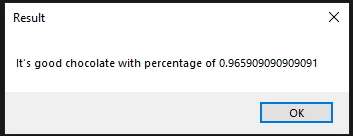
Za unos parametara napravljena je forma (slika 6.1.) u kojoj se unose potrebne vrijednosti, te se pritiskom da tipku „Predict“, podaci šalju web servisu na obradu. Dobiveni rezultat ispisuje dobivenu vrijednost za klasu čokolade. Ispis loše, dobre i izvrsne kvalitete čokolade vidljiv je na slikama 6.2 - 6.4.



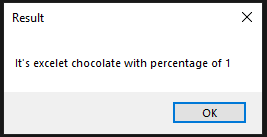
**Slika 6.1. –** Izgled forme za unos podataka



**Slika 6.2. –** Izgled dobivenog rezultata kada je u pitanju loša čokolada



**Slika 6.3. –** Izgled dobivenog rezultata kada je u pitanju dobra čokolada



**Slika 6.4. –** Izgled dobivenog rezultata kada je u pitanju izvrsna čokolada

# ZAKLJUČAK

Strojno učenje, iako veoma star postupak obrade informacija sve je popularniji u modernom vremenu. Postoje mnoge mogućnosti i primjene, a jedna od njih je opisana u ovom radu. Koristeći skup podataka sa popularnog repozitorija Kaggle, napravljen je i istreniran model koji na temelju ulaznih parametara vrši klasifikaciju čokolade.

Na temelju istreniranog modela izrađeni su web servis i aplikacija koja ga koristi. Aplikacija u vrlo brzom roku može odrediti rezultata za testirane parametre i vjerojatnost tog rezultata sa veoma velikom točnosti.

Aplikacija bi mogla imati nekoliko primjena, a najočitija je da gurmanima čokolade pomogne u pri klasifikaciji čokolade.

# LITERATURA

1. [https://www.kaggle.com/rtatman/chocolate-bar-ratings](https://www.kaggle.com/rtatman/chocolate-bar-ratings%20)
2. <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-decision-forest>
3. <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-decision-jungle>
4. <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-neural-network>
5. <https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-logistic-regression>

# POVEZNICE

1. <https://github.com/szabbo/ruapProject>