Mendelova univerzita v Brně
Provozně ekonomická fakulta

KVALITA BÍLÉHO VÍNA

Neuronové sítě v aplikacích – LS 2016

Členové týmu:

Bc. David Krénar, Bc. Petr Sadovský

Brno 2016

Obsah

1 Úvod a cíl práce

1.1 Úvod

Kvalita vína je důležitým prvkem pro obchod. Vyhodnocování kvalit vín zabraňuje nelegálnímu falšování vín a přispívá ke zvyšování kvality vína na trhu pro zákazníka.

Víno lze klasifikovat lidskými experty nebo fyzikálně-chemickými laboratorními testy – hodnoty pH (pH), podíl alkoholu (alcohol) či stanovení hustoty (density).

1.2 Cíl práce

Cílem práce je predikovat (klasifikovat víno do skupin) kvalitu vína na základě vstupních proměnných. S tímto cílem jsou spojeny dílčí cíle, mezi něž patří zobrazení relevantních vstupních trénovacích dat, vytvořit neuronovou síť a naučit tuto síť pomocí **backpropagation** (backward propagation of errors) algoritmu rozlišovat jednotlivá vína na základě vstupů a určit chybu (MCE – Misclassification Error) učení.

2 Data

2.1 O datasetu

Dataset je volně k dispozici pro vědecké účely po odcitování zdroje:

```
P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
Modeling wine preferences by data mining from physicochemical proper-
ties.

In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016
                [Pre-press (pdf)]
http://www3.dsi.uminho.pt/pcortez/winequality09.pdf
                [bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib
```

Pro náš příklad jsme vybraly pouze data obsahující vzorky *bílého* vína. Informace o kyselinách, pH, alkoholu, hustotě, atd. Tyto informace získané z fyzikálně-chemického měření jsou vstupními daty datasetu a výstupními informacemi jsou hodnoty získané z mediánu alespoň ze tří hodnocení provedených vinařskými odborníky.

Dataset bílého vína je tvořen portugalským vínem „Vinho Verde“. Z důvodu ochrany osobních údajů jsou k dispozici pouze fyzikálně-chemické vstupy a smyslové proměnné jako výstupy (např. nejsou zde zahrnuty informace jako typ hroznů, značka vína, tržní cena vína, atd.).

Rozdělení vín do tříd není rovnoměrné (data obsahují více normálních vín než skvělých či chudých).

Data obsahují 4898 instancí bílého vína a každá tato instance má 11 atributů (vlastností) + výstupní atribut udávající do jaké třídy dané víno spadá.

Některé atributy mají mezi sebou vztah (spojení), tudíž dává smysl použít funkci výběru na některé vlastnosti (atributy) vzorků při zobrazení jejich vzájemných vlastností.

2.2 Atributy

Vstupní proměnné (na základě fyzikálně-chemických testů):

1. fixed acidity (tartaric acid – g / dm^3) – většina kyselin obsažených ve víně, které se snadno neodpařují (především kyselina vinná)
2. volatile acidity (acetic acid – g / dm^3) – množství kyseliny octové ve víně (příliš vysoká úroveň této kyseliny vede k nepříjemné chuti vína a připomíná spíše chuť octu)
3. citric acid (g / dm^3) – nalezena v malých množstvích ve víně; kyselina citronová přidává „čerstvost“ k chuti vína
4. residual sugar (g / dm^3) – zbylé množství cukru po fermentaci; je vzácné najít vína s méně než 1 g / litr vína a s více než 45 g / litr vína.

Tabulka 1: Rozdělení vín podle obsahu zbytkového cukru

Množství zbytkového cukru (gram / litr)	Druh vína
Více než 45	Sladká
12 – 45	Polosladká
4 – 12	Polosuchá
Méně než 4	Suchá

5. chlorides (sodium chloride – g / dm^3) – množství soli ve víně
6. free sulfur dioxide (mg / dm^3) – množství volného oxidu siřičitého (SO2) – volná forma existuje v rovnováze mezi molekulárním oxidem siřičitým (rozpuštěný plyn) a iontem hydrogensiřičitanu ($HSO3^-$); zabraňuje růstu mikroorganismů a oxidaci vína
7. total sulfur dioxide (mg / dm^3) – celkové množství SO2 – množství volných i vázaných forem SO2 v nízkých koncentracích; SO2 je většinou nezjistitelný ve víně, ale koncentrace volného SO2 větší než 50 ppm (parts per milion – „částic na milion) se projevuje na chuti vína a je patrné při přičichávání k vínu
8. density (g / cm^3) – hustota vody je v závislosti množství procent alkoholu a cukru ve vodě
9. pH – popisuje, jak kyselé či zásadité víno je na stupnici od 0 (velmi kyselé) do 14 (velmi jednoduché); většina vín je mezi 3 – 4 na stupnici pH
10. sulphates (potassium sulphate – g / dm3) – sírany jsou aditiva vína, které mohou přispět k úrovni oxidu siřičitého (SO2), který působí jako antibakteriální a antioxidační prvek
11. alcohol (% by volume) – procentuální obsah alkoholu ve víně

Výstupní proměnná (na základě smyslových dat od odborníků vín):

12. quality (score between 0 (very bad) and 10 (excellent)) – hodnota vína udělena odborníky vín (bílá vína mezi 3 – 9)

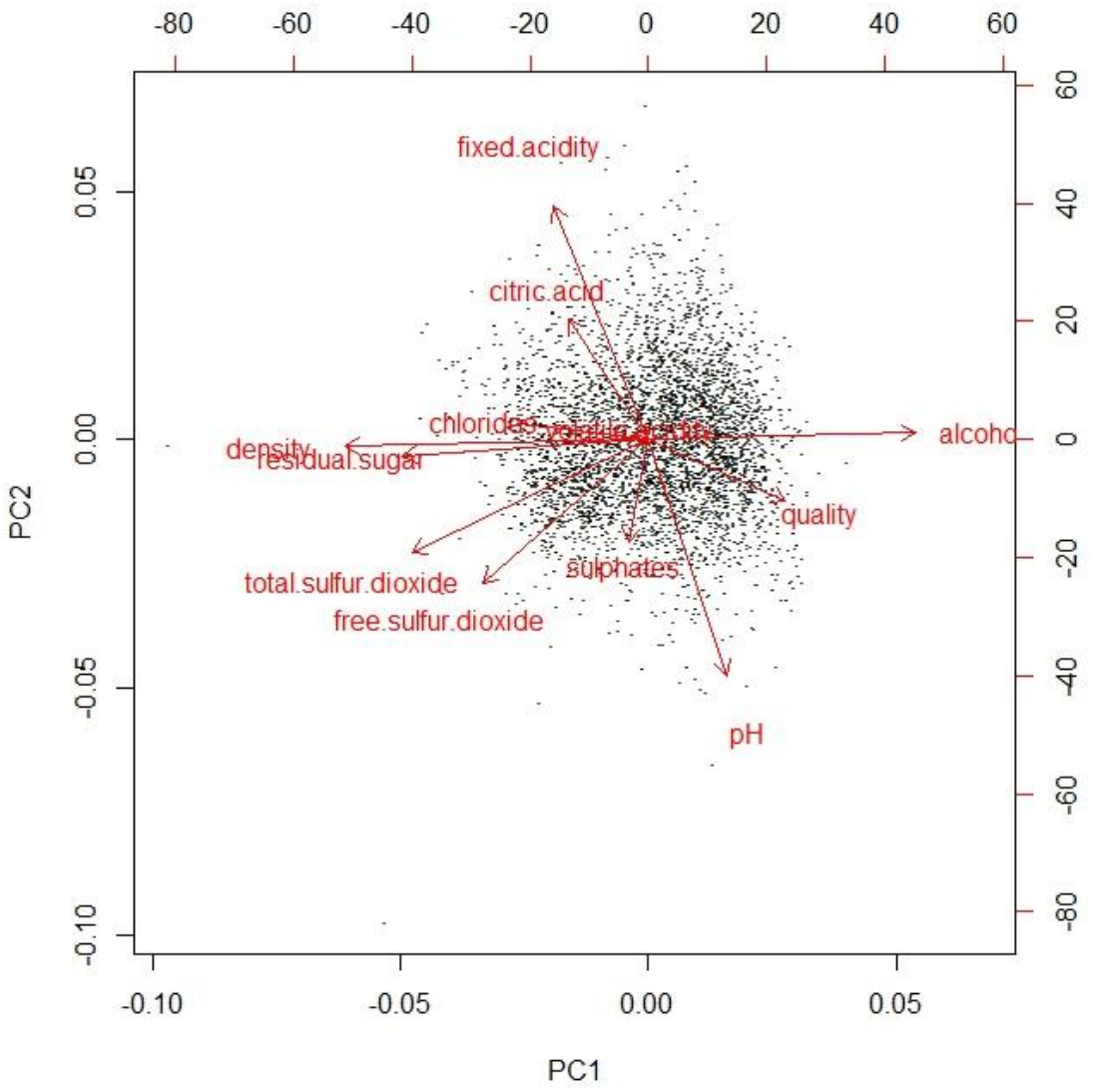
Tabulka 2. Ukazuje základní sumarizovaná vstupní data datasetu bílého vína.

Atributy (jednotky)	**min**	**mean**	**max**
Fixed acidity (g / dm3)	3,8	6,8	14,2
Volatile acidity (g /dm3)	0,08	0,26	1,1
Citric acid (g/dm3)	0	0,32	1,66
Residual sugar (g / dm3)	0,6	5,2	65,8
Chlorides (g / dm3)	0,009	0,043	0,346
Free sulfur dioxide (mg/dm3)	2	34	289
Total sulfur dioxide (mg/dm3)	9	134	440
Density (g/cm3)	0,9871	0,9937	1,039
pH	2,72	3,18	3,82
Sulphates (g / dm3)	0,22	0,47	1,08
Alcohol (% vol.)	8,0	10,4	14,2

3 Metodika

3.1 Analýza hlavních komponent pro bílé víno (Principal Component Analysis)

Abychom mohly s daty pracovat, musíme si je nejdříve ukázat a zjistit, které vlastnosti vína víno nejvíce ovlivňují a zda mezi nimi není nějaký bližší vztah. K tomuto se dá využít analýza hlavních komponent (PCA), která nám ukáže rozmístění datových bodů v grafu.



Obrázek 1: Biplot: symboly odpovídají jednotlivým datovým bodům promítnutým do roviny

Na tomto grafu je zajímavé to, že podle jeho dvou hlavních komponent, je kvalita vína do značné míry ve vztahu s obsahem alkoholu a množstvím kyseliny vinné (fixed acidity), či velikostí pH (odvíjí se od množství kyselin ve víně).

Nicméně tyto dvě hlavní vlastnosti se dají snadno získat i pro jiná vína a proto mají praktický význam. To platí především pro obsah alkoholu, který je uveden na každé vinětě láhve vína. Obsah kyseliny vinné lze jednoduše změřit pH metrem.

Ale pozor, pouze na základě těchto dvou znalostí nelze předpovídat kvalitu vína, tyto znalosti nám poskytnout pouze nějaké vodítku při výběru vína.

Tento graf PCA byl proveden v programu R (v. 3.3.0). Níže uveden část převzatého zdrojového kódu[1].

```
data_file = "winequality-white.csv"

wine <- read.csv( data_file, sep=';', header = TRUE )

numel = length( as.matrix( wine )) / length( wine )

pcx <- prcomp( wine, scale = TRUE )
biplot( pcx, xlabs = rep( '.', numel ))
```

3.2 Terminologie

V této části uvedeme kategorie vín podle hodnocení odborníků.

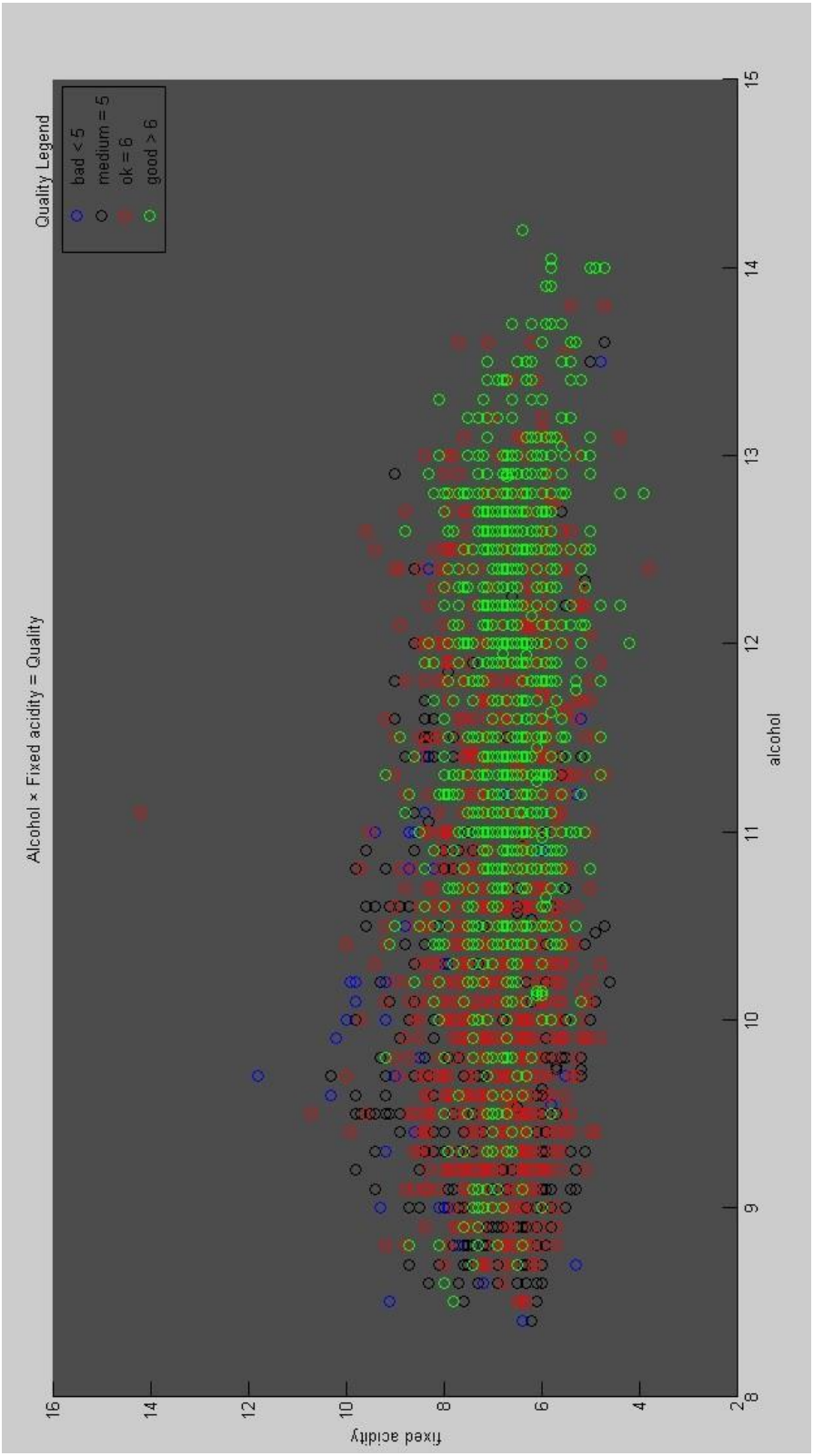
- *Bad* – špatná vína mají hodnocení 4 a méně
- *Medium* – podprůměrná vína mají hodnocení rovno 5
- *Ok* – ucházející vína mají hodnocení rovno 6 (převládají)
- *Good* – dobrá vína mají hodnocení 7 a více

Z grafu níže (viz Obrázek 2) lze odvodit pár zajímavostí. Pokud obsah alkoholu je aspoň 11 % či více máme kvalitní víno. Na druhou stranu pokud je obsah alkoholu menší než 10 % nemusí se jednat zrovna o kvalitní víno.

To není vše, pokud je obsah kyselosti menší, tím je víno lepší. Tudíž můžeme mít kvalitní víno s nižším obsahem alkoholu a malou kyselostí (vína v rozmezí mezi 10 a 12 na ose alkoholu).

Z tohoto vyplývá, že souzení vína pouze pomocí dvou hlavních komponent je poněkud zjednodušené, ale aspoň jsme získaly představu, které z daných vlastností z datasetu jsou ty hlavní. Samozřejmě existují i jiné faktory,, které by bylo potřeba zvažovat při výběru kvalitních vín, jako například věk. Většina má raději starší, uležené víno i když se najdou i milovníci mladých vín (jednoroční).

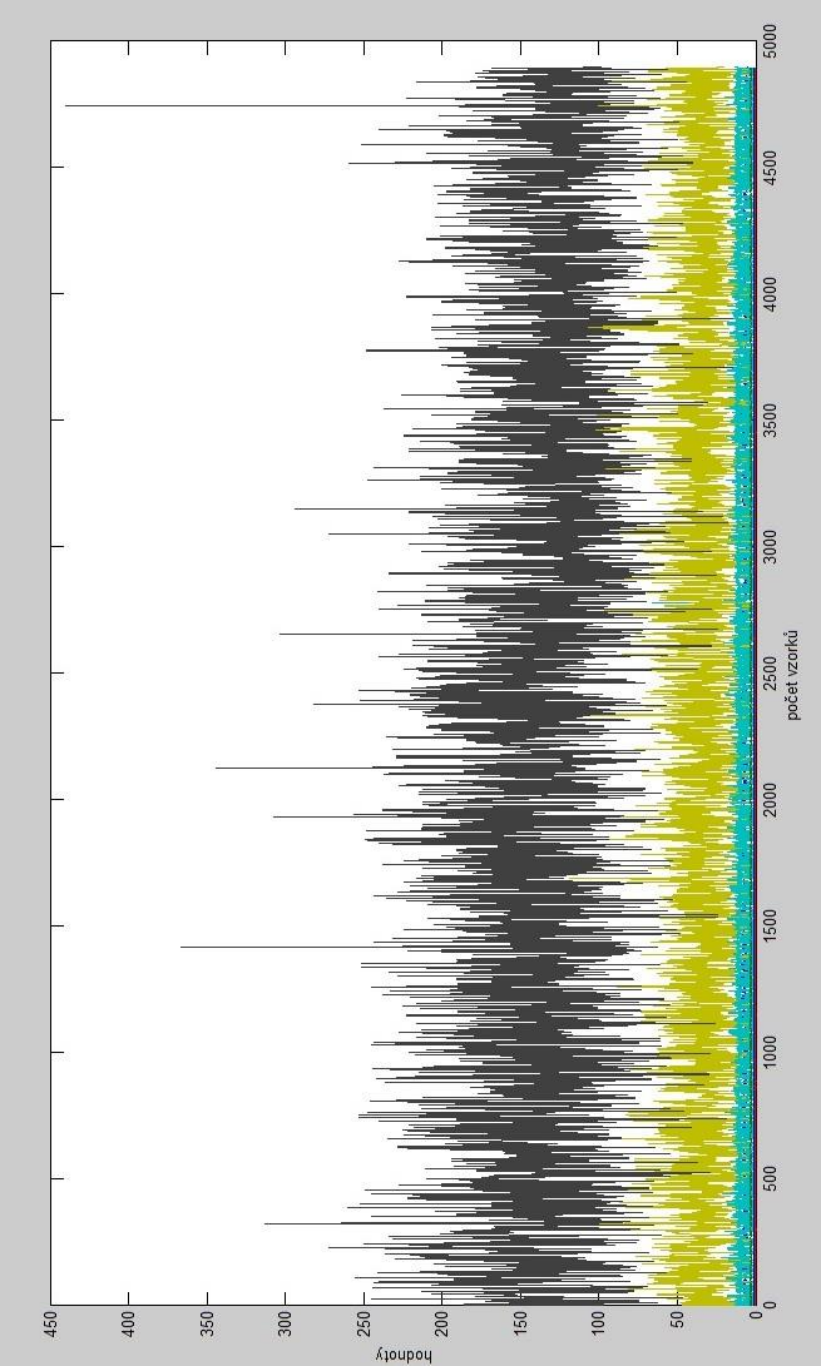
[1] Zdroj: https://github.com/zygmuntz/wine-quality/blob/master/pca_red.r



Obrázek 2: Graf rozložení kvality vín podle obsahu alkoholu a množství kyseliny vinné (fixed acidity)

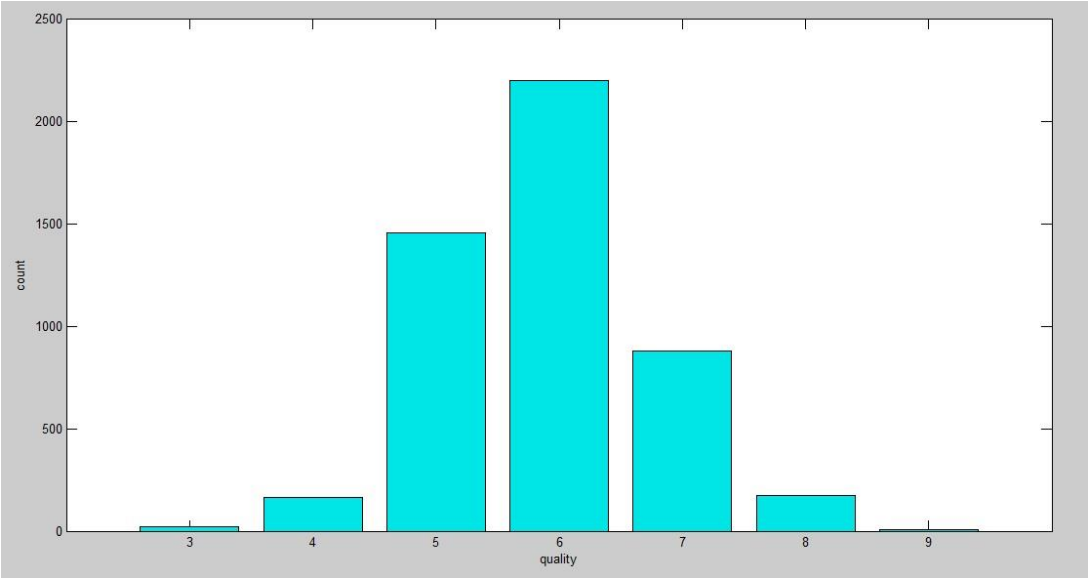
4 Matlab

Vykreslení vstupních hodnot.



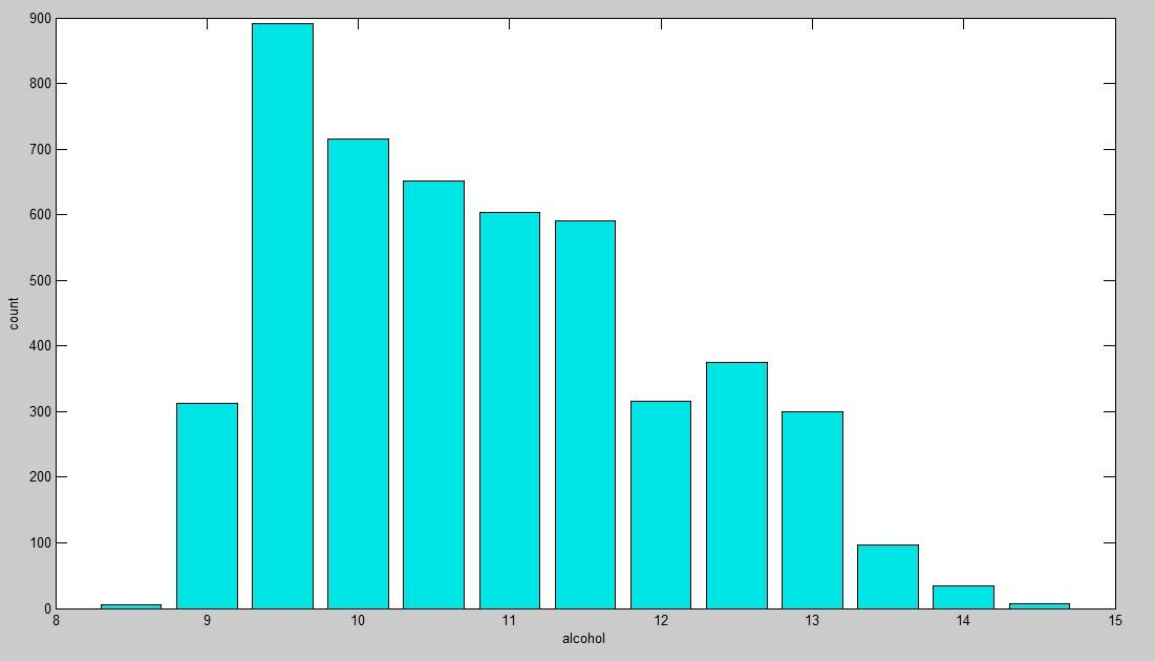
Obrázek 3: Nenormalizované vzorky bílého vína. Hodnoty mezi 0 a 440.

Pohled na množství vín podle jejich kvality. Z grafu je patrné, že největší zastoupení v datasetu mají vína ucházející před podprůměrnými.

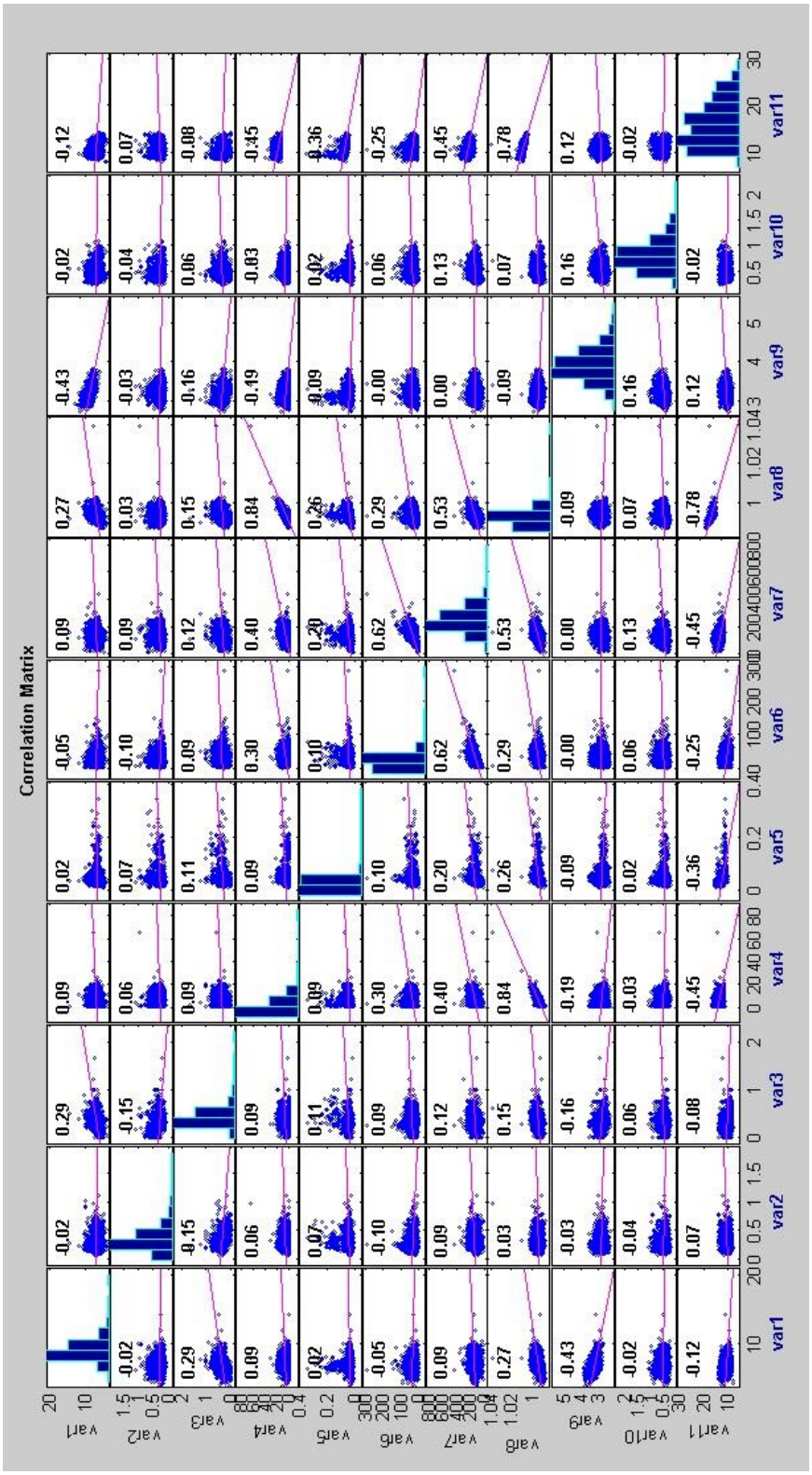


Obrázek 4: Graf množství vín podle jejich kvality

Pohled na množství vín dle obsahu alkoholu. Z je patrné, že největší zastoupení mají vína s 9,5 % a 10 % obsahu alkoholu v jednom litru vína.



Obrázek 5: Graf množství vín dle obsahu alkoholu



Obrázek 6: Korelační matice vstupních dat (zobrazuje vzájemný vztah vstupních veličin)

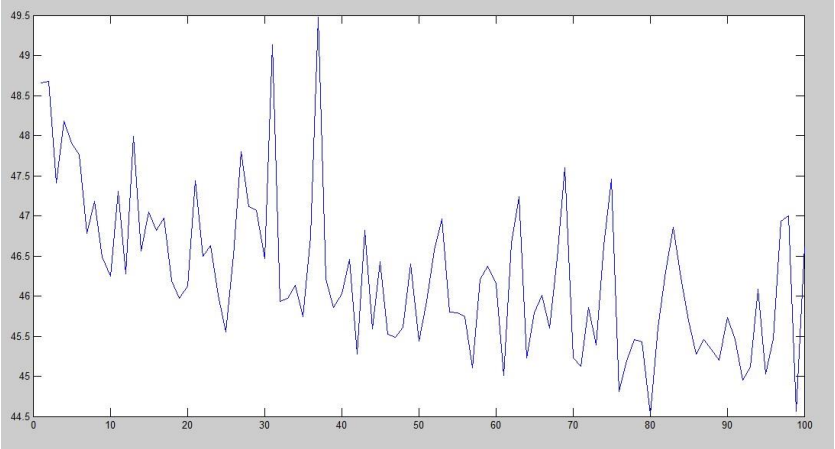
Legenda k obrázku (Obrázek 6):

- var1 – fixed acidity
- var2 – volatile acidity
- var3 – citric acid
- var4 – residual sugar
- var5 – chlorides
- var6 – free sugar dioxide
- var7 – total sugar dioxide
- var8 – density
- var9 – pH
- var10 – sulphates
- var11 – alcohol

4.1 Trénování a testování

Pro trénování neuronové sítě je třeba určit počet iterací (dobu) učení, počet neuronů ve skryté vrstvě a hodnotu (míru) učení.

1. **Trénování:**
    - hidden_neurons (počet neuronů ve skryté vrstvě) = 10
    - iterations (doba učení) = 100
    - learn_rate (míra učení) = 0.1

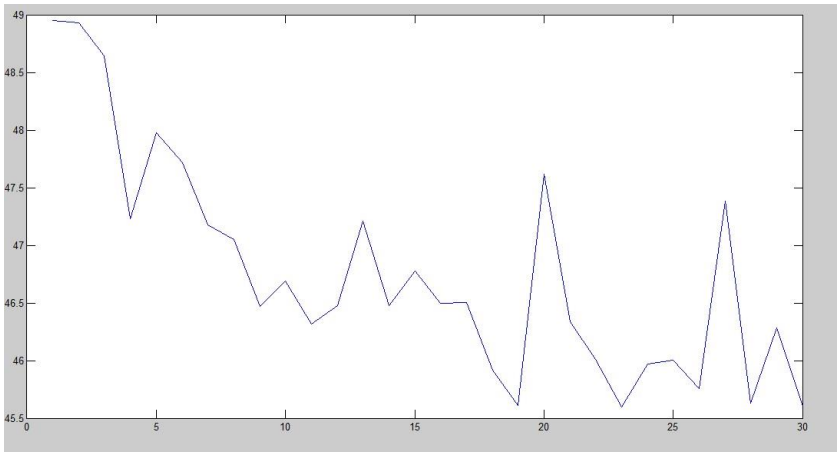


Obrázek 7: Průběh učení 1. trénování

Chyba učení: 9,667 %
Správně kvalifikovaných: 90,233 %

2. **Trénování**

- hidden_neurons (počet neuronů ve skryté vrstvě) = 10
- iterations (doba učení) = 30
- learn_rate (míra učení) = 0.1

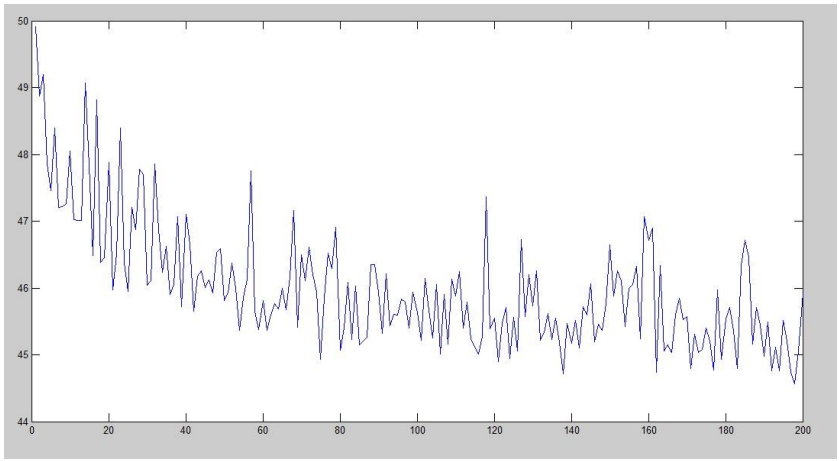


Obrázek 8: Průběh učení 2. trénování

Chyba učení: 9,8942 %
Správně kvalifikovaných: 90,1058 %

3. **Trénování**

- hidden_neurons (počet neuronů ve skryté vrstvě) = 10
- iterations (doba učení) = 200
- learn_rate (míra učení) = 0.1

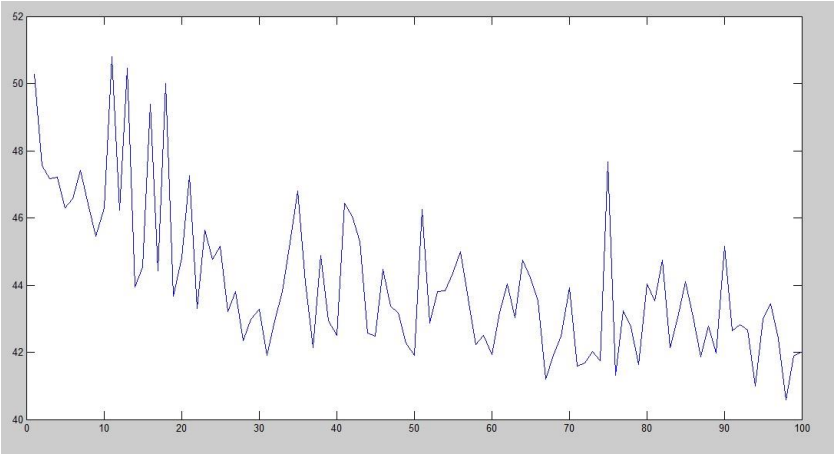


Obrázek 9: Průběh učení 3. trénování

Chyba učení: 10,0485 %
Správně kvalifikovaných: 89,9515 %

4. **Trénování**

- hidden_neurons (počet neuronů ve skryté vrstvě) = 25
- iterations (doba učení) = 100
- learn_rate (míra učení) = 0.1

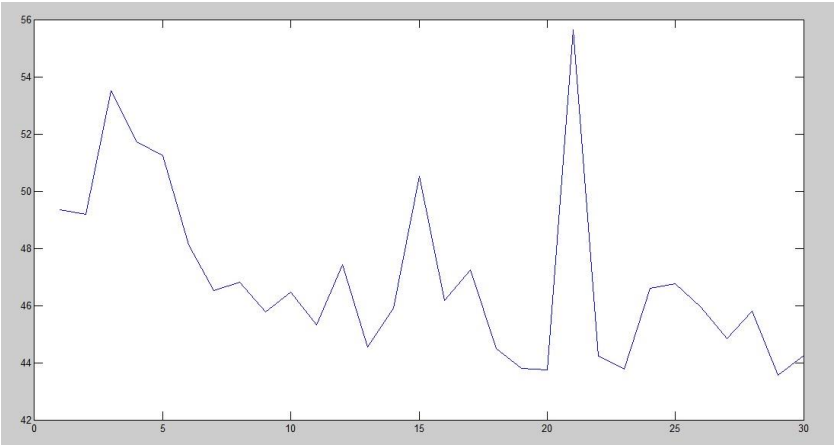


Obrázek 10: Průběh učení 4. trénování

Chyba učení: 9,0940 %
Správně kvalifikovaných: 90,9060 %

5. **Trénování**

- hidden_neurons (počet neuronů ve skryté vrstvě) = 25
- iterations (doba učení) = 30
- learn_rate (míra učení) = 0.1

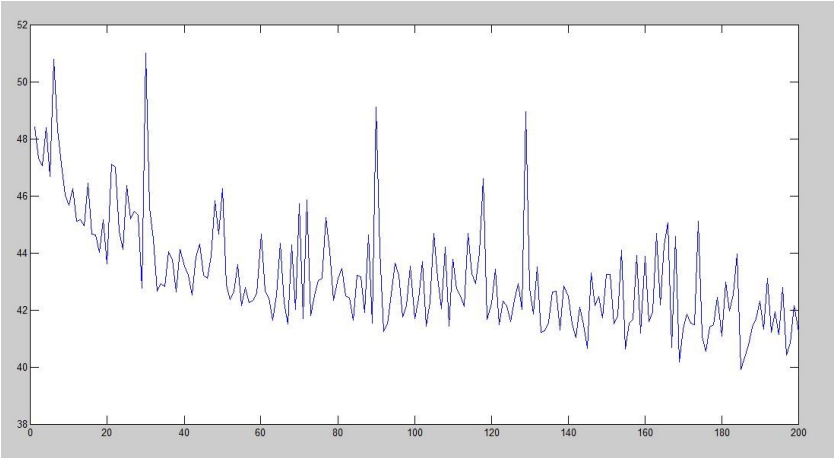


Obrázek 11: Průběh učení 5. trénování

Chyba učení: 9,5865 %
Správně kvalifikovaných: 90,4135 %

6. Trénování

- hidden_neurons (počet neuronů ve skryté vrstvě) = 25
- iterations (doba učení) = 200
- learn_rate (míra učení) = 0.1

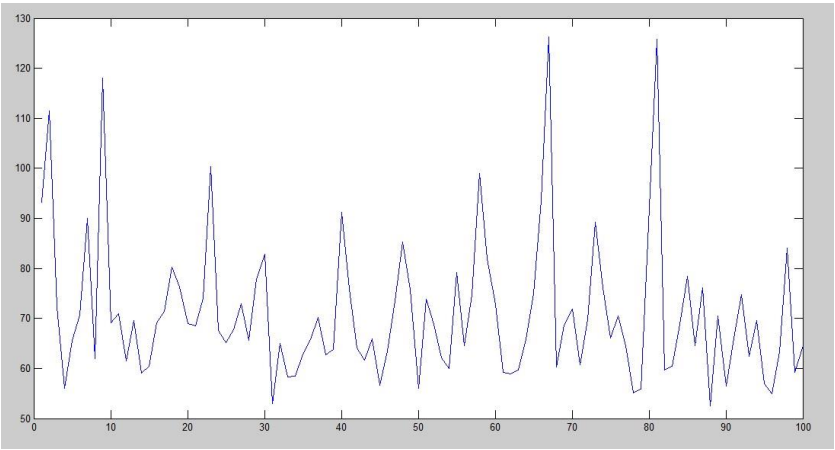


Obrázek 12: Průběh učení 6. trénování

Chyba učení: 8,9667 %
Správně kvalifikovaných: 91,0333 %

7. Trénování

- hidden_neurons (počet neuronů ve skryté vrstvě) = 10
- iterations (doba učení) = 100
- learn_rate (míra učení) = 1

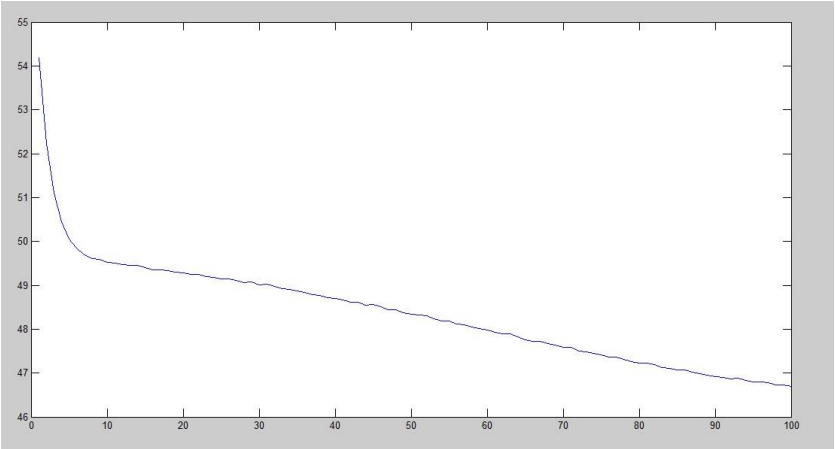


Obrázek 13: Průběh učení 7. trénování

Chyba učení: 10,0485 %
Správně kvalifikovaných: 89,9515 %

8. Trénování

- hidden_neurons (počet neuronů ve skryté vrstvě) = 10
- iterations (doba učení) = 100
- learn_rate (míra učení) = 0.001



Obrázek 14: Průběh učení 8. trénování

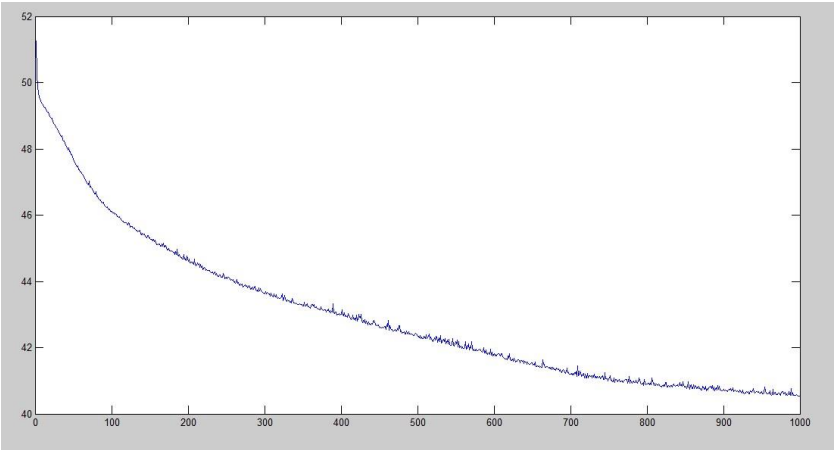
Chyba učení: 10,0508 %
Správně kvalifikovaných: 89,9492 %

5 Závěr

Z grafů lze vyčíst, že lepší výsledky dostáváme přidáváním neuronů do skryté vrstvy, zvyšováním doby učení a snižováním míry učení (lepší průběh učení).

9. Trénování

- hidden_neurons (počet neuronů ve skryté vrstvě) = 25
- iterations (doba učení) = 1000
- learn_rate (míra učení) = 0.001



Obrázek 15: Průběh učení 9. (posledního) trénování

Chyba učení: 8,6913 %
Správně kvalifikovaných: 91,3087 %

Tabulka 2: Výsledky po trénování neuronové sítě (zlomek dat)

Trénovací hodnota	Predikovaná hodnota	Rozdíl trénovací a predikované hodnoty (%)	Procentuální rozdíl trénovací a predikované hodnoty (%)
6.0000	6.0787	0.0787	1.3111
6.0000	6.0383	0.0383	0.6376
6.0000	6.0632	0.0632	1.0530
6.0000	5.9421	-0.0579	0.9657
6.0000	6.1029	0.1029	1.7154
7.0000	7.0442	0.0442	0.6311
6.0000	6.1950	0.1950	3.2495
6.0000	5.6655	-0.3345	5.5742