Wine Quality predikcia

Nikola Kulikova

January 13, 2023

Abstrakt

V tomto dokumente sa budem venovať predikcií kvality vína na základe jeho vlastnosti ako sú percento alkoholu, pH, hustota a podobne. Na túto predikciu budem využívať strojové učenie.

V ďalších častiach sa budem venovať opisu problému, ktorý budem riešiť, dáta, s ktorými budem pracovať. Ďalej opíšem prístupy, ktoré som použila na riešenie problému a ich výsledky. Kod je dostupny na https://github.com/nikolakulikova/WineQuality

Opis problému

V tomto projekte riešim klasifikáciu dát na základe viacerých parametrov. Moje dáta sú jednotlivé vína spolu s ich parametrami opísanými v časti o datasete. Pomocou klasifikátora následne budem môcť predikovať kvalitu vín. Okrem predikcie sa budem taktiež venovať hľadaniu dôležitých parametrov z datasetu a vizualizácia dát.

Dataset

Na tento projekt som si našla dataset na Kaggle (https://www.kaggle.com/datasets/yasserh/wine-quality-dataset).

Tento dataset obsahuje vína s ich parametrami. Víno je špecifikované 11 paramaterami, konkrétne - pevná kyslosť, prchavá kyslosť, kyselina citrónová, zvyškový cukor, chloridy, voľný oxid siričitý, celkový oxid siričitý, hustota, pH, sírany, podiel alkoholu. Každé víno má taktiež určenú svoju kvalitu v škále od 0 po 10 (0- najhoršie, 10-najlepšie). Všetky atribúty sú zadané ako reálne čísla a teda nebude treba ich predspracovávať.

Dataset neobsahuje okrajové príklady a to buď veľmi kvalitné alebo veľmi nekvalitné vína. Väčšina príkladov sa nachádza v strede škály alebo v jej blízkosti. Prvotne som si myslela, že dataset je dostatočne veľký ale neskôr sa ukázalo, že by sa zišlo viacej dát.

Prístupy a výsledky

V nasledujúcej časti sa budem venovať podrobnejšie prístupom, ktoré som si na riešenie môjho problému vybrala. Rovnako sa budem venovať ich výsledkom a následným porovnaním s baseline, ktorá vracia najčastejšiu kvalitu vína (v mojom prípade 5). Data som si ako prvé rozdelila na trénovacie a testovacie. Trénovacie budem používať na natrénovanie klasifikárova a testovaciu na ohodnotenie výsledkov.

Support Vector Classification

Ako prvý prístup som použila Support Vector Classification(SVC), ktorý sme už predtým využívali na cvičeniach. Pre tento klasifikátor viem nastaviť niekoľko parametrov, podľa ktorých ho viem napasovať presnejšie na môj problém. Pri hľadaní týchto parametrov som využila funkciu get_hyper_param_SVC vytvorenú na to. Ako najlepšie parametre vyšli C=0.2 a gamma=0.16. Na obrázku 1 je zobrazený priebeh učenia môjho modelu pri rôznych veľkostiach trénovacej a testovacej množiny.

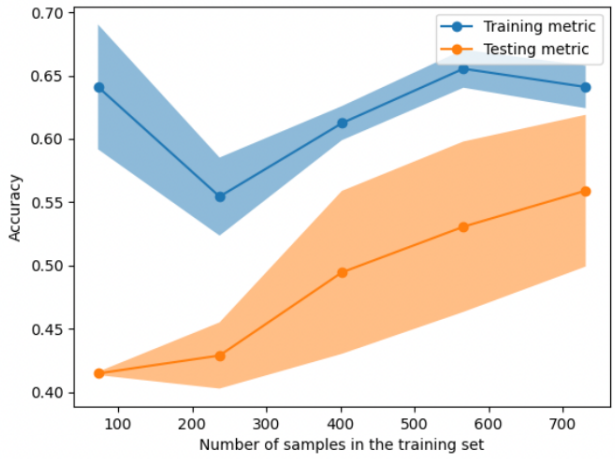

Figure 1: Priebeh učenia SVC

Random Forest Classifier

Ďalším prístupom , ktorý som výskušala bol Random Forest Classifier(RFC). S týmto prístupom som sa už stretla na cvičeniach, kde som taktiež zistila, že tento klasifikátor má dva atribúty a to počet estimátorov a maximálnu hĺbku lesu. Opäť pri hľadaní týchto parametrov som využila funkciu get_hyper_param_RFC vytvorenú na to. Ako najlepšie parametre vyšli n_estimators=64 a max_depth=ý. Na obrázku 2 je zobrazený priebeh učenia môjho modelu pri rôznych veľkostiach trénovacej a testovacej množiny.

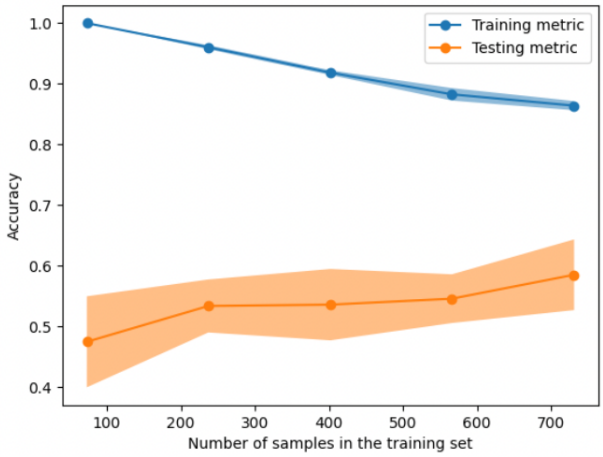

Figure 2: Priebeh učenia Random Forest Classifier

K Neighbors Classifier

Posledným klasifikátorom, ktorý som použila je K-Neighbors Classifier(KNC). Tento klasifikátor sme na cvičeniach nemali. Tento klasifikátor dostáva atribút počet susedov - k, ktorý slúži na určenie počtu susedov na základe ktorých sa bude klasifikátor rozhodovať, ktorú triedu u priradí. Priradí sa tá trieda, ktorá je najčastejšia medzi jeho k-najbližšími susedmi. Rovnako ako aj pri ostatných klasifikátoroch som si vytvorila funkciu na nájdenie najlepšieho počtu susedov. V mojom prípade to vyšlo na 67 susedov. Priebeh učenia tohto klasifikátora môžeme vidieť ako pri ostaných na obrázku 3.

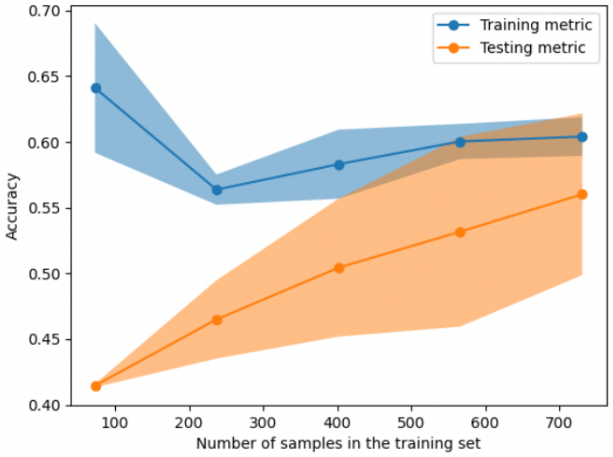

Figure 3: Priebeh učenia K Neighbors Classifier

Multilayer perceptrons

Ako posledný prístup som vyskúšala Multilayer perceptrons(MLP), ktorý sme taktiež mali na cvičeniach. V tomto prístupe je možné zvoliť si počet skrytých vrstiev a ich aktivačné funkcie. Ďalej je možne zvoliť si rýchlosť učenia, počet epochov, ktoré má môj model na samotné učenie. Určite som nedosiahla najlepšiu kombináciu týchto parametrov nakoľko je to časovo náročne. Po vyskúšaní viacerých možností som vybrala tú, ktorá vracala najlepšie výsledky. A teda moja konštrukcia má jednu vstupnú vrstvu s aktivačnou funkciou sigmoid a jednou skrytou vrstvou, ktorá obsahuje taktiež funkciu sigmoid. Konštrukcia má jednu výstupnú vrstvu s funkciou softmax. Počet epochov je 100 a rýchlosť učenia modelu je nastavená na 0.001. Ako optimizér som použila Adam.

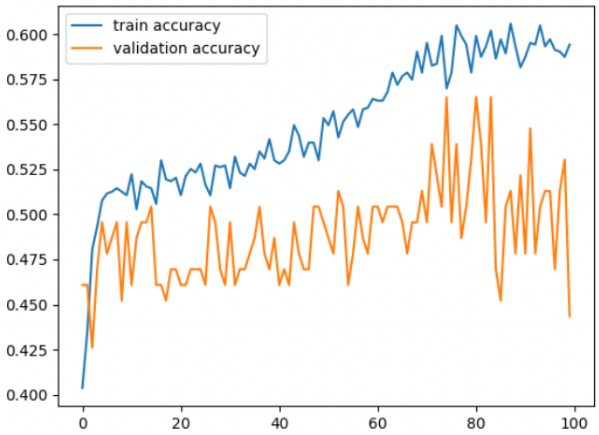

Figure 4: Priebeh učenia MLP

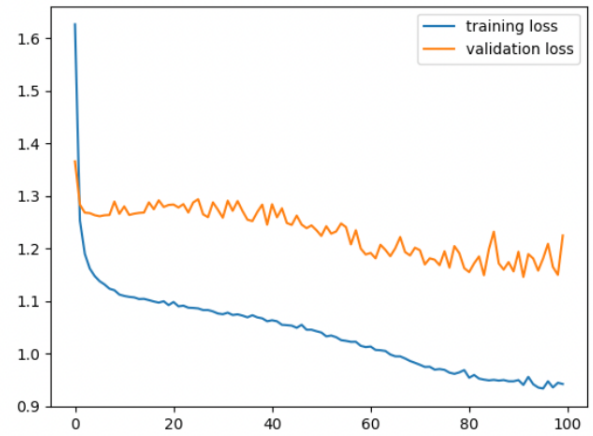

Figure 5: Loss MLP

Porovnanie:

	baseline	SVC	RFT	KNC	MLP
train	0.414	0.644	0.839	0.603	0.596
test	0.454	0.646	0.633	0.641	

Hľadanie dôležitých parametrov

Lasso

Tento prístup sme mali spomenutý na hodine. Pri použití metódy Lasso je dôležité nájsť vhodnú α. Tú som hľadala while cyklom až kým nebola nulová a v medzikrokoch som si zapamätala tú najlepšiu, pre ktorú bol cross validačné skóre najmenšie. Nájdené najdôležitejšie parametre teda sú: fixed acidity, volatile acidity, free sulfur dioxide, total sulfur dioxide, sulphates, alcohol.

Vizualizácia dát

Z nájdených najdôležitejších parametrov som si vybrala 3 najdôležitejšie z nich (alcohol, sulphates, volatile acidity) a následne zobrazila ich vzájomnú závislosť:

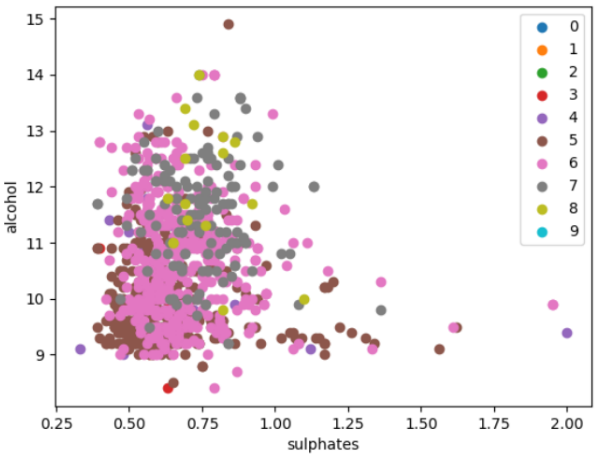

Figure 6: Vizualizácia sulfátov a percent alkoholu

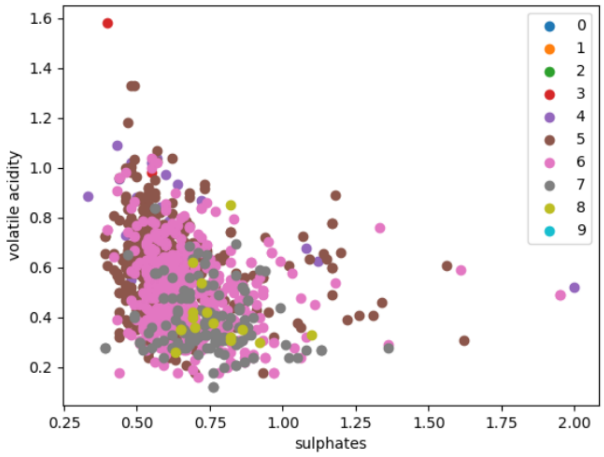

Figure 7: Vizualizácia sulfátov a prchavej kyslosti

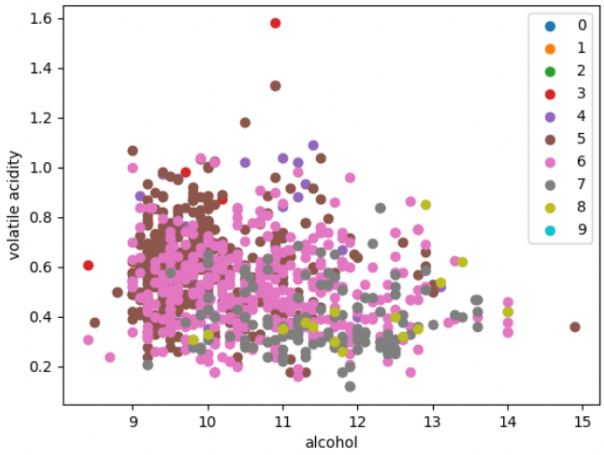

Figure 8: Vizualizácia percent alkoholu a prchavej kyslosti

Záver

Cieľom môjho projektu bolo predikovať kvalitu vína podľa jeho atribútov. Na tento projekt som si našla dataset, ktorý po vyskúšaní viacerých prístupov sa ukázal ako nie úplne dostatočne veľký. Ako najdôležitejšie parametre pre predikciu som vybrala fixed acidity, volatile acidity, free sulfur dioxide, total sulfur dioxide, sulphates, alcohol pomocou metódy Lasso a alcohol, sulphates, volatile acidity, citric acid, chlorides, density pomocou metódy information gain. Za

pomoci najdôležitejších atribútov som si vybrala 3 z nich a dataset aj vizualizovala. Na riešenie problému klasifikácie som použila rôzne klasifikátory a multi layer perceptom. Najlepším riešením vyšiel K Neighbors Classifier a SVC s podobnou presnosťou. Možným vylepšením by mohlo byť vyskúšanie väčšieho datasetu.