**Project 5 – Advanced Data Mining Applications**

**CS548 / BCB503 Knowledge Discovery and Data Mining - Fall 2017**

**Prof. Carolina Ruiz**

**Student: Tianyu Wu**

|  |  |
| --- | --- |
| Description of the particular problem within the selected data mining topic to be addressed in this project | /15 |
| Description of the approach used in this project to tackle the above problem.  *All data mining techniques you use in this project for pre-processing, mining and evaluation must have been covered in class during this semester.* | /25 |
| Description of the dataset selected | /15 |
| Appropriateness of the dataset selected with respect to this topic/problem | /10 |
| Guiding questions | /10 |
| Preprocessing | /10 |
| **Experiments:**   * Sufficient & coherent | /25 |
| * Objectives, Data, Additional Pre/Post-processing | /20 |
| * Presentation of results | /20 |
| * Analysis of results | /30 |
| Overall discussion, comparisons, and conclusions | /20 |
| TOTAL | /200 |

Total Written Report: \_\_\_\_\_\_\_\_\_\_\_\_\_\_/200 = \_\_\_\_\_\_\_\_\_\_\_/100

Class Presentation: \_\_\_\_\_\_\_\_\_\_\_/100

Class participation during project presentation: \_\_\_\_\_\_\_\_\_\_\_/100

*Do not exceed the given page limits for this written report*

**Topic: Text Mining <at most 1 page>**

1. **Description of the particular problem within the selected data mining topic to be addressed in this project:**

Datasets sometimes contains large amount of texts information that hard to KDD, even though human can read and understand them. Also, some datasets contains texts that would take long time for human reading or translating which is not efficient for human resources. So text mining techniques are necessary for this scenario to make conclusions or find patterns by machines.

1. **Description of the approach used in this project to tackle the above problem:**

In this dataset, the descriptions were human sommeliers reports which describe wines taste, look, feel, etc, but the machine cannot understand or analyse. I used vectorizer technique with tf-idf term weighting in this project, to find best description of varieties of wines that machine can understand and analyse, also to find some other interesting patterns by using classification in this dataset.

1. **Dataset Name: Wine Reviews**
2. **Where found: Kaggle**
3. **Dataset Description:**

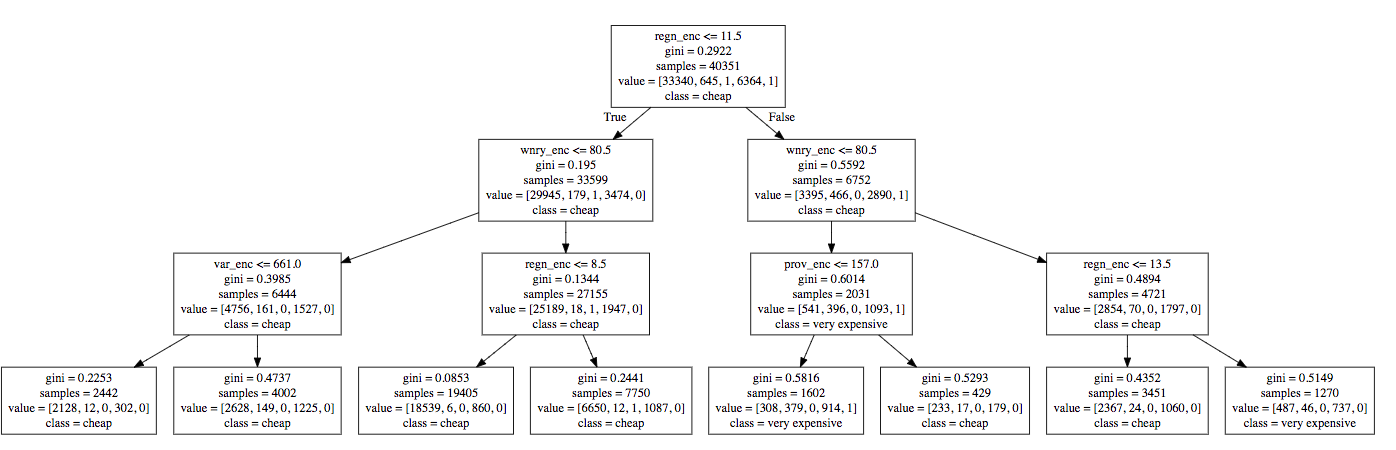
This wine reviews dataset was scraped from *WineEnthusiast* during the week of June 15th 2017, it contained 150930 instances along with 10 attributes. The data consisted of 10 fields: ***Country***(the wine is made from), ***Description***(from sommeliers describing taste, smell, look, feel, etc), ***Designation***(where the grapes are from), ***Points***(rated on a scale of 1-100), ***Price***(cost of one bottle), ***Province***(province or state that made the wine), ***Region\_1***(wine growing area), ***Region\_2***(more specific region), ***Variety***(the type of grapes), and ***Winery***(winey made the wine).

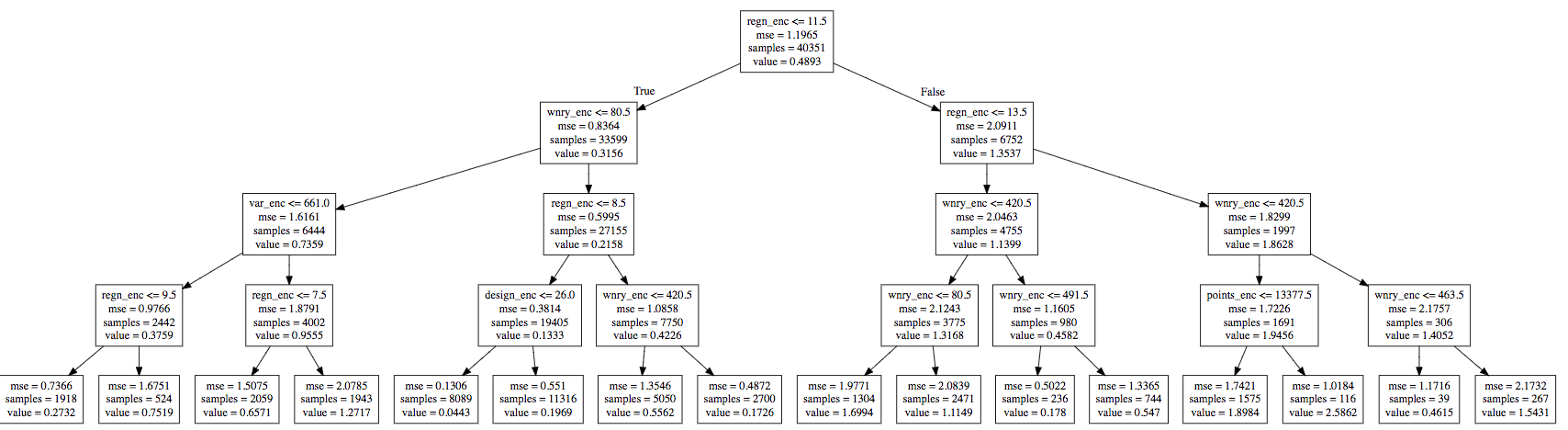
1. **Initial data preprocessing, if any:**

Missing value only appears in attribute ***Designation***(30.37%), ***Price***(9.07%), ***Province***(0.78%), ***Region\_1***(16.60%), and ***Winery***(0.02%), after removing missing value and duplicate instances, I finally got 97156 instances along with 9 attributes, namely, ***Country, Description, Designation, Points, Price, Province, Region\_1, Variety, and Winery.*** I tried to use online translating APIs to translate French, Italian, and Spain, because those language were not encoded as Ascii, and only converted them into UTF-8 would break the words. Unfortunately, popular translator APIs were not available currently, so I had to manually remove all those strange foreign characters first then stemmed them. I also used *LabelEncoder()* from sklearn.preprocessing to transform nominal strings into values in the classification parts, because Decision Tree and Regression Tree technique in Python can only handle value based attributes. At last, I used *inverse\_transform()* to get the encoding regulations then transfer the value back to strings.

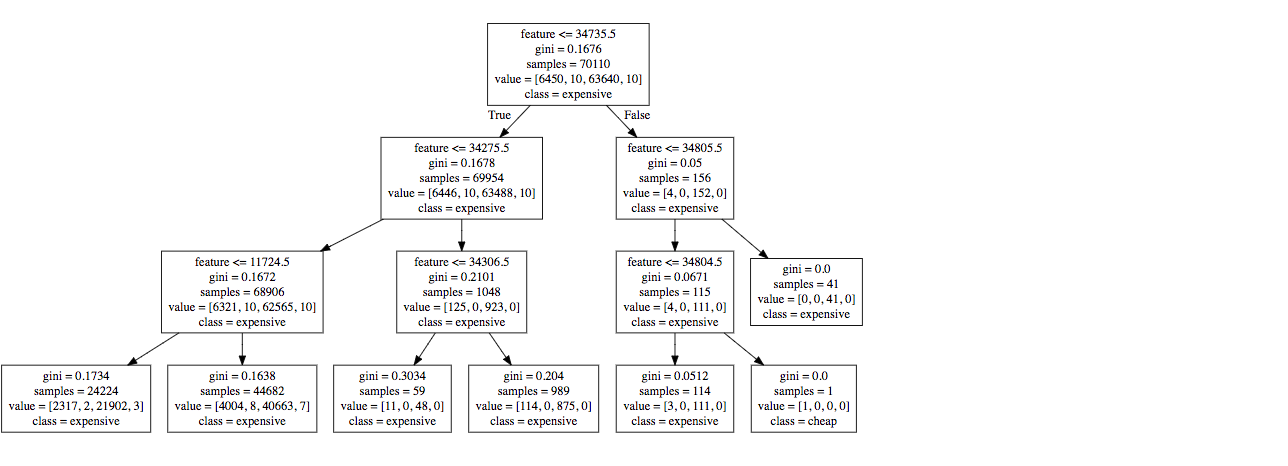
1. **Three Guiding Questions about the dataset domain:**
2. What are the best description features of different varieties of wine?
3. What are the best description features of different country’s wine?
4. Can we predict wine price range based on description features?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Summary of Experiments.** *At most 2 page.* | | | | | | |
| **Tool** | **Pre-process** | **Mining**  **Technique** | **Results**  **Variety: Description Features**  **or Tree structure** | **Time**  **taken** | **Evaluation**  **TF-IDF score:** | **Observations about experiment**  **Observations about visualization**  **Interpretation results** |
| Python | Stemmed and punctuation | Text mining | **Chardonnay:***green tannic*  **Cabernet Sauvignon:** *superoak, featureless, unforgiv, raspberryflavor*  **Red Blend:***sugar dessert*  **Pinot Noir:** *medicine flavor* | 115s | [0.886]  [0.785, 0.696, 0.660, 0.658]  [0.644]  [0.600] | Stemming and punctuation leaded to huge different results, that’s because foreign language are not encoded into Ascii, and stemmed them will get unreadable characters. In this experiment, we could observe such that Chardonnay taste like green tannic and Red Blend more was suitable for dessert, also Pinot Noir taste like medicine. |
| Python | No | Text mining | **Cabernet Sauvignon:** *green tannic*  **Red Blend:**est, *coarseness, unforgiving*  **Chardonnay:** *sugar dessert,sweet sugar*  **Pinot Noir:***slight berry,oddly,medicinal flavors* | 127s | [0.886]  [0.661,0.588,0.583]  [0.629,0.602]  [0.580,0.574,0.568] |
| Python | Stemmed and punctuation | Text mining | **Italy:***green tannic,light rose,unforgiv*  **France:***superoak,dilut tart*  **Spain:***featureless,promis fruit,way toast*  **US:***raspberryflavor,sugar dessert,cherryberri spice* | 237s | [0.886,0.839,0.660]  [0.786,0.657]  [0.696,0.688,0.565]  [0.658,0.643,0.642] | The reason that leaded into two different results was the same as above. We could observe the results of this experiment such like Italian like to make rose and tannic flavor wine, Spanish like to make toast, cherry flavor, and American like to make barrel, cherryberry, and dessert wine. |
| Python | No | Text mining | **Italy:***green tannic,light rose,promise fruit*  **US:***wine company, est,sugar dessert,barrel*  **France:***slight berry,medicinal flavor,viognier tropical*  **Spain:***rosemount,syrupy cherry,juice tannins* | 245s | [0.886,0.840,0.676]  [0.691,0.661,0.629,0.614]  [0.580,0.568,0.556 ]  [0.527,0.481,0.480] |
| Python | Encoded nominal attributes into value-based label except description | Decision tree | 0 [label="regn\_enc <= 11.5\ngini = 0.2922\nsamples = 40351\nvalue = [33340, 645, 1, 6364, 1]\nclass = cheap"] ; | 0.557s | cross\_val\_score: 0.847 | Graph view result will be introduced in result analysis. Based on cross validation score and error matrix, we could say using Region, Winery, Designation, and Points to predict wine price could be pretty accurately. Although, encoding huge amount of nominal attributes into value label by LabelEncoder() and for reversing operation by inverse\_transform() in Python seems clumsy, WEKA would be much more convenient. |
| Python | Encoded nominal attributes into value-based label except description | Regression Tree | 0 -> 1 [labeldistance=2.5, labelangle=45, headlabel="True"] ;  2 [label="varieti\_enc <= 661.0\nmse = 1.6161\nsamples = 6444\nvalue = 0.7359"] ; | 0.101s | PCC: 0.830  MAE: 0.255  MSE: 0.371 |
| Python | Ranked wine price into cheat  ,normal,expensive,very expensive and extreme expensive.  Encoded nominal attributes into value-based label. Cut out redundant data to make price distribution balanced. | Text mining and Classification | 0 [label="feature <= 14274.5\ngini = 0.4622\nsamples = 1000\nvalue = [680, 50, 270]\nclass = cheap"] ;  1 [label="feature <= 13613.0\ngini = 0.5067\nsamples = 237\nvalue = [145, 11, 81]\nclass = cheap"] ;  0 -> 1 [labeldistance=2.5, labelangle=45, headlabel="True"] ; | 35.26s | cross\_val\_score: 0.826 | The score was surprisingly high, which means we could use description features to predict wine price range. We could observe the results in next section, that there are particular words to describe wines in different price range. Such like cheap wine is sweet fruity and overripe, extreme expensive wine is intense elegant, upfront delicious and soft decad. |

***Decision Tree using other attributes except description features, Fig[1]***



***Regression Tree using other attributes except description features, Fig[2]***



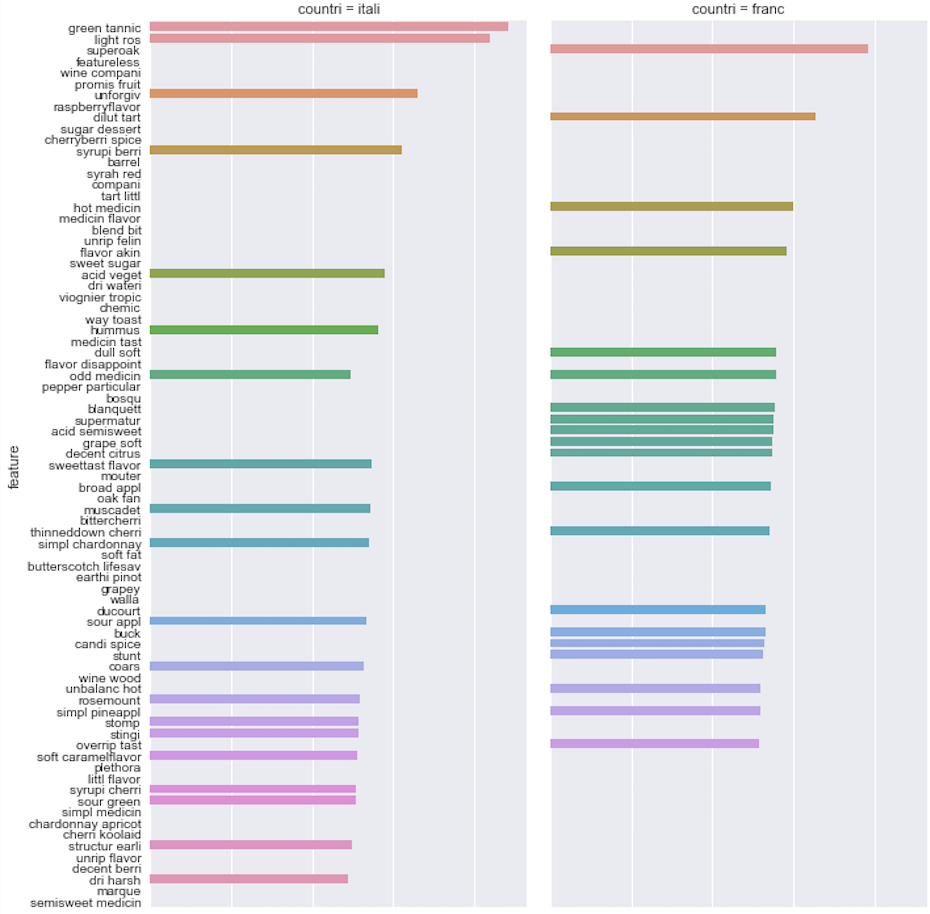
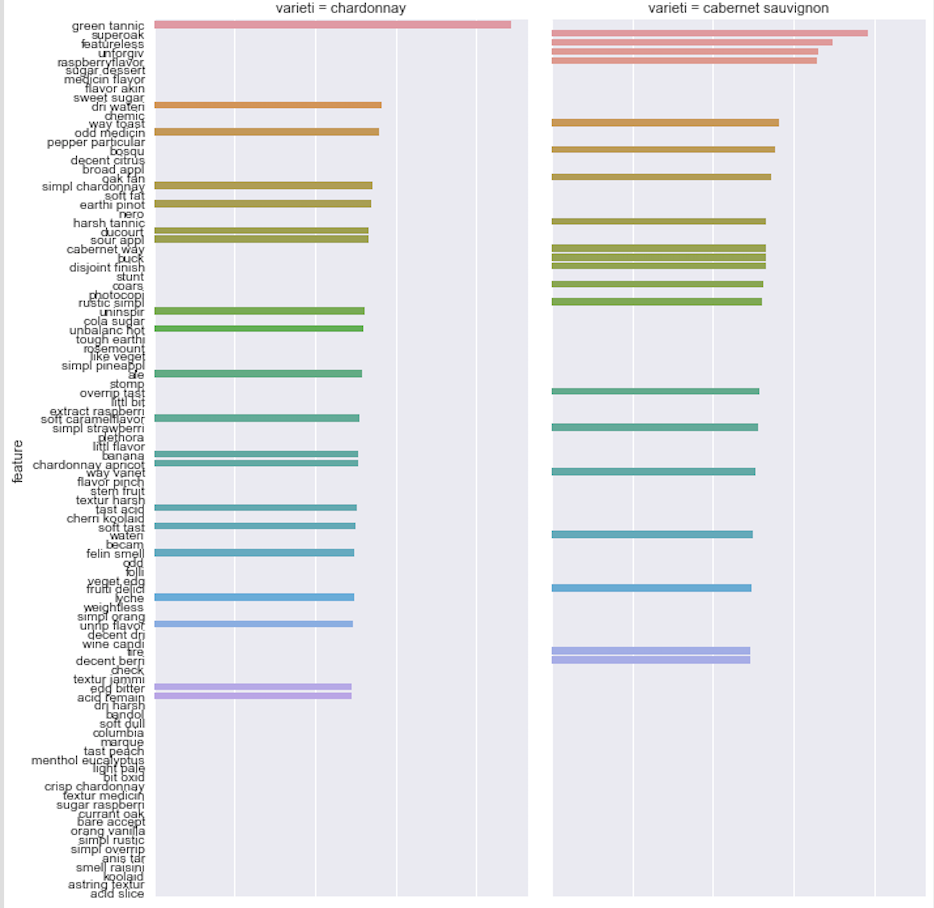
***Decision Tree using description features, Fig[3]***

Each nominal attribute was encoded into value-based data, the encoding regulations were according to LabelEncoder() and can be inversed by inverse\_transform() in sklearn.preprocessing package.

**Analysis of Results: (at most 1 page)** 1. Analyze the effect of varying parameters/experimental settings on the results. 2. Analyze the results from the point of view of the Domain, and discuss the answers that the experiments provided to your guiding questions. 3. Include and explain (some of) the best / most interesting results you obtained in your experiments. 4. Include visualizations.

Stem and punctuation helped text mining technique with non-ASCII data and foreign characters, this was an significant variance step in preprocessing that could impact results. Using Decision Tree for classification made better results than Regression Tree, because only one attribute ***Price*** was continuous data in this dataset. By experiments, I got the following interesting results that I would never know without the experiments: the features of each variety of wines, the features of each country’s wines, and we could predict wine’s price range by sommeliers’ descriptions. These are some details:

Red Blend wine was coarseness and suitable by dessert. Pinot Noir wine tasted like slight berry, oddly, and medicinal flavors(with pretty high TF-IDF scores). US liked to make berry and sweet flavor wine while Italy liked to make tart wine. Spain wine were always criticized as featureless and way toasted. Besides that, we could use decision tree to classify or even predict wine’s price based on its attributes or descriptions. Based on Fig[1], we could get price of wine by if its region, winery, variety and province. Based on Fig[3], we could classified wine by its descriptions like abund fresh, accent anis. For further understanding, I will add all attributes label encoding into appendix. Classification results are included in the previous page.



***Feature extraction by variety in TF\_IDF score Feature extraction by country in TF\_IDF score***

**Summary of what you learned in this project:**

The first knowledge I’ve learned from this project is that text mining technique spends a lot of time, not only summarize what the domain and natural language knowledge we should let machine recognize, but also the time of running the codes(took much longer than previous projects). Second, text mining has to be used along with other data mining techniques such like classification, clustering, association, etc. If not, the results of text mining cannot be used in many scenarios. Third, text mining is such an useful and interesting technique that could let machine know what we are talking, reporting and documenting. Future techknowledge could be more like letting machine know what we are doing, not letting us know what machines are doing. Last but not the least, teamwork could create much more interesting ideas than myself, I miss the time that I worked with my teammates!

**Appendix**

***Price Encoder:*** 0 : cheap 1 : expensive 2 : extreme expensive 3 : normal 4 : very expensive

***Points Encoder:*** 0 : 80 1 : 81 2 : 82 3 : 83 4 : 84 5 : 85 6 : 86 7 : 87 8 : 88 9 : 89 10 : 90 11 : 91 12 : 92 13 : 93 14 : 94 15 : 95 16 : 96 17 : 97 18 : 98 19 : 99 20 : 100

***Winery Encoder:*** 1 : son 2 : 10 knot 3 : 1000 stori 4 : 1040fu 5 : 1070 green 6 : 10span 9 : 12c wine 10 : 14 hand 11 : 16x20 12 : 1789 wine 13 : 181 15 : 1850 19 : 2 cocki sister 21 : 21 gram 22 : 2820 wine co 23 : 2hawk 24 : 2nd chanc 25 : 2plank 27 : 3 hors ranch vineyard 29 : 3 spell 30 : 3 steve wineri 34 : 31st state 35 : 32 wind 36 : 39 37 : 3cv 38 : 3fool 39 : 4 bear 40 : 401k 41 : 428 wine 45 : 5 point cellar 46 : 50 harvest 47 : 6 north 48 : 60 north 49 : 6th sens 50 : 7 heaven chard 52 : 75 wine co 53 : 868 estat 58 : aaron 59 : abacela 66 : abandon 74 : abbey creek 77 : abbeyvill 81 : abeja 83 : aberr cellar 84 : abiou 85 : abiqua wind 88 : acacia 94 : acker pond 95 : ackerman 100 : acorn …(13715)

***Variety Encoder:*** 1 : aglianico 5 : albario 8 : aleatico 11 : alicant bouschet 12 : aligot 15 : alvarelho 18 : angevin 21 : appl 26 : arnei 32 : auxerroi 35 : baco noir 38 : barbera 41 : black monukka 42 : black muscat 46 : blaufrnkisch 51 : bordeauxstyl red blend 52 : bordeauxstyl white blend 58 : cabernet 59 : cabernet blend 60 : cabernet franc 63 : cabernet francmalbec 64 : cabernet francmerlot 66 : cabernet merlot 68 : cabernet pfeffer 69 : cabernet sauvignon 72 : cabernet sauvignoncabernet franc 73 : cabernet sauvignoncarmenr 74 : cabernet sauvignonmalbec 75 : cabernet sauvignonmerlot 77 : cabernet sauvignonsangioves 78 : cabernet sauvignonshiraz 79 : cabernet sauvignonsyrah 82 : cabernetsyrah 86 : carignan 87 : carignangrenach 91 : carmenr 100 : cayuga …(615)

***Region Encoder:*** 1 : aglianico 5 : albario 8 : aleatico 11 : alicant bouschet 12 : aligot 15 : alvarelho 18 : angevin 21 : appl 26 : arnei 32 : auxerroi 35 : baco noir 38 : barbera 41 : black monukka 42 : black muscat 46 : blaufrnkisch 51 : bordeauxstyl red blend 52 : bordeauxstyl white blend 58 : cabernet 59 : cabernet blend 60 : cabernet franc 63 : cabernet francmalbec 64 : cabernet francmerlot 66 : cabernet merlot 68 : cabernet pfeffer 69 : cabernet sauvignon 72 : cabernet sauvignoncabernet franc 73 : cabernet sauvignoncarmenr 74 : cabernet sauvignonmalbec 75 : cabernet sauvignonmerlot 77 : cabernet sauvignonsangioves 78 : cabernet sauvignonshiraz 79 : cabernet sauvignonsyrah 82 : cabernetsyrah 86 : carignan 87 : carignangrenach 91 : carmenr 100 : cayuga …(1179)

***Designation Encoder:*** 1 : red tabl wine 2 : vineyard 3 : 0 degre 4 : 0 degre dri 6 : 1 11 : 1 liter 13 : 10 14 : 10 acr 17 : 10 vine age seri 23 : 10 year old vine 24 : 10 year tawni 25 : 100 musqu clone 27 : 100 skin ferment 28 : 100 year old vine 31 : 1000 vine 35 : 10th anniversari reserv 41 : 11 43 : 1105 46 : 1149 ros 47 : 115667 50 : 12 reserv 60 : 1470 estat 70 : 15th anniversari ottimo red wine 73 : 16 row 80 : 1762 81 : 1772 82 : 1772 barrel select 83 : 1772 edna ranch 86 : 181 88 : 181merlot 93 : 1861 vineyard 97 : 1866 reserv 98 : 1869 99 : 1870 …(27262)

***Province Encoder:*** 1 : red tabl wine 2 : vineyard 3 : 0 degre 4 : 0 degre dri 6 : 1 11 : 1 liter 13 : 10 14 : 10 acr 17 : 10 vine age seri 23 : 10 year old vine 24 : 10 year tawni 25 : 100 musqu clone 27 : 100 skin ferment 28 : 100 year old vine 31 : 1000 vine 35 : 10th anniversari reserv 41 : 11 43 : 1105 46 : 1149 ros 47 : 115667 50 : 12 reserv 60 : 1470 estat 70 : 15th anniversari ottimo red wine 73 : 16 row 80 : 1762 81 : 1772 82 : 1772 barrel select 83 : 1772 edna ranch 86 : 181 88 : 181merlot 93 : 1861 vineyard 97 : 1866 reserv 98 : 1869 99 : 1870

***Description Features Encoder:*** 0 : abacela 1 : abandon 2 : abbey 3 : abbott 4 : abbrevi 5 : abbrevi finish 6 : abeja 7 : abil 8 : abil age 9 : abil grow 10 : abil produc 11 : abil ripen 12 : abl 13 : abl age 14 : abl handl 15 : abound 16 : abound dri 17 : abras 18 : abrupt 19 : absenc 20 : absenc fruit 21 : absenc oak 22 : absenc rich 23 : absenc structur 24 : absolut 25 : absolut beauti 26 : absolut decant 27 : absolut delici 28 : absolut dri 29 : absolut dryness 30 : absolut firstrat 31 : absolut gorgeous 32 : absolut love 33 : absolut stun 34 : absorb 35 : absorb new 36 : abund 37 : abund black 38 : abund cherri 39 : abund coffe 40 : abund flavor 41 : abund floral 42 : abund fresh 43 : abund fruit 44 : abund note 45 : abund peach 46 : abund red 47 : abund ripe 48 : abund stone 49 : abund sweet 50 : abund vanilla 51 : abv 52 : acacia 53 : acacia barrel 54 : acacia flower 55 : accent 56 : accent acid 57 : accent add 58 : accent anis 59 : accent aroma 60 : accent balanc 61 : accent barrel 62 : accent black 63 : accent blackberri 64 : accent bodi 65 : accent bright 66 : accent brisk 67 : accent butter 68 : accent caramel 69 : accent cedar 70 : accent cherri 71 : accent cinnamon 72 : accent citrus 73 : accent clove 74 : accent cocoa 75 : accent coffe 76 : accent cola 77 : accent continu 78 : accent core 79 : accent cranberri 80 : accent crisp 81 : accent dark 82 : accent dash 83 : accent deep 84 : accent dri 85 : accent dusti 86 : accent earth 87 : accent earthi 88 : accent fennel 89 : accent fine 90 : accent finish 91 : accent firm 92 : accent flavor 93 : accent floral 94 : accent fresh 95 : accent fruit 96 : accent green 97 : accent herb 98 : accent hint 99 : accent honey 100 : accent intrigu …(57825)