U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

ELEMENTS OF ARTIFICAL INTELLIGENCE AND DATA SCIENCE

2022-23

PRACTICAL WORK / ASSIGNMENT TWO

DATA EXPLORATION AND ENRICHMENT FOR SUPERVISED CLASSIFICATION

PORTUGUESE WINE THROUGH THE LENSES OF VIVINO'S USERS

TEACHERS: LUÍS PAULO REIS, PEDRO FERREIRA, DAVID APARÍCIO

STUDENT: VÍTOR BRUNO DANTAS RAMALHOSA FERREIRA (201109428)

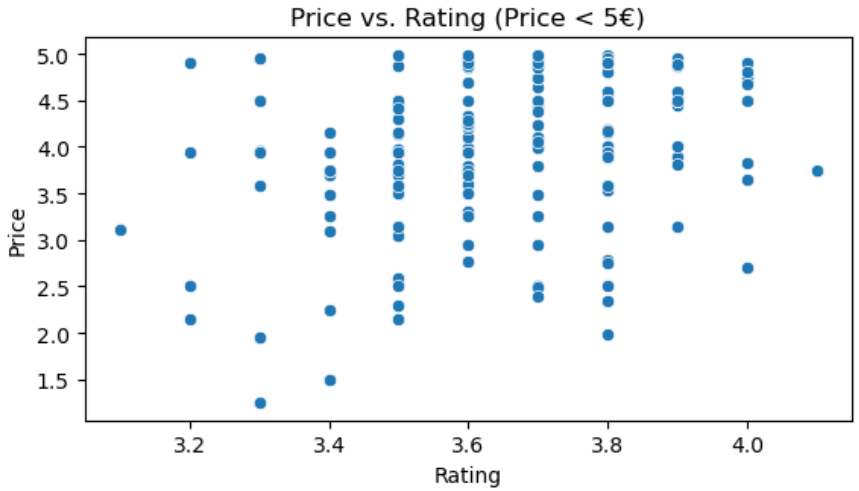
DATA GATHERING AND PRE-PROCESSING

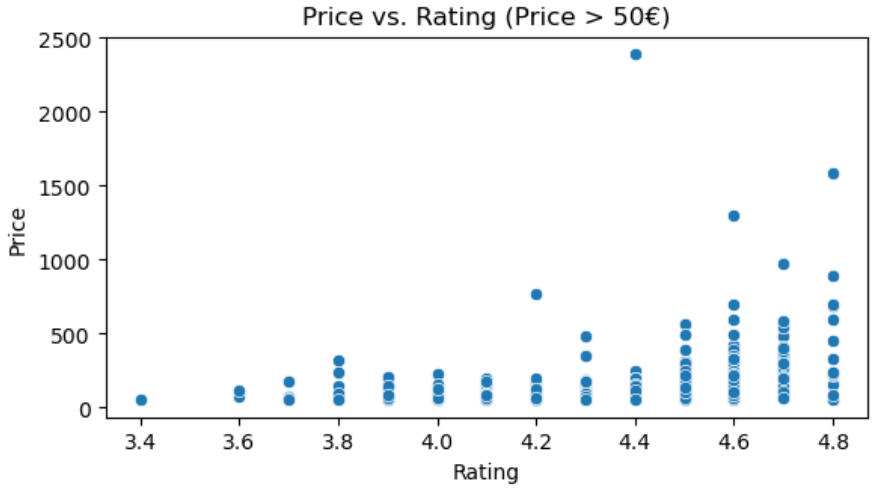
- Portuguese wine data was gathered using Vivino Webscraper (the code was adapted from the Vivino Webscraper developed in the project: https://www.kaggle.com/datasets/joshuakalobbowles/vivino-wine-data). Data was collected on the following types of wines: **red**, **white**, **rosé**, **sparkling**, **dessert**, **fortified**. Each reference displayed the following information: 1) Winery / Producer; 2) Year / Vintage; 3) Wine Id.; 4) Wine Name; 5) Rating (possible range: 0.5 to 5.0); 6) Number of Reviews; 7) Price; 8) Country; 9) Region. Data was cleaned by performing the following tasks: 1) removing duplicate entries; removing entries with less than 50 ratings, uniformizing data (e.g. region “Alentejano” was converted to “Alentejo”; region “Douro Superior” (subregion of Douro) was converted to the general region (“Douro”)); 2) a column that contains information about the wine type (Red, White, Rosé, Sparkling, Dessert, Fortified) was added; 3) the datasets were merged into a single one (pt_wine_merged.csv), resulting in **3281 wine references**.

	Winery	Year	Wine ID	Wine	Type	Rating	Num_Reviews	Price	Country	Region
0	São João	1980	76447	Porta dos Cavaleiros Dão Tinto 1980	Red	4.3	81	37.50	Portugal	Dão
1	São João	1985	76446	Frei João Reserva 1985	Red	4.0	75	28.00	Portugal	Bairrada
2	São João	1985	1689617	Porta dos Cavaleiros Dão Reserva Tinto 1985	Red	4.2	57	29.90	Portugal	Dão
3	Casa Ferreirinha	1989	75980	Reserva Douro 1989	Red	4.5	157	180.00	Portugal	Douro
4	Cabriz	1990	1148718	Dão Reserva 1990	Red	3.8	525	39.95	Portugal	Dão
...	...	...	...	...	...	...	...	...	...	...
3276	Vasques de Carvalho	N.V.	3882596	40 Years Old Tawny Port N.V.	Fortified	4.6	164	206.00	Portugal	Porto
3277	Pacheca	N.V.	4904195	40 Years Tawny Porto N.V.	Fortified	4.6	637	100.80	Portugal	Porto
3278	Rozès	N.V.	1507187	Over 40 Years Old Tawny Port N.V.	Fortified	4.6	301	185.00	Portugal	Porto
3279	Vieira de Sousa	N.V.	3942014	Very Old White Port N.V.	Fortified	4.6	50	214.81	Portugal	Porto
3280	Taylor's	N.V.	6250054	50 Year Old Tawny Port N.V.	Fortified	4.7	91	300.00	Portugal	Porto

DATA ANALYSIS AND CONCLUSIONS

- The mean, maximum and minimum of the variables Rating, Number of Reviews and Price were calculated:
  - Wine Rating: 2.70 (Lowest), 3.99 (Average), 4.80 (Highest);
  - Number of Reviews: 50 (Lowest), 419 (Average), 86421 (Highest);
  - Price:1.25€ (Lowest), 36.39€ (Average), 2390.84€ (Highest).
- The average rating and price of wines were compared based on different price categories (less than 5€, between 5€ and 10€, between 11€ and 15€, between 16€ and 25€, between 25€ and 50€, more than 50€):





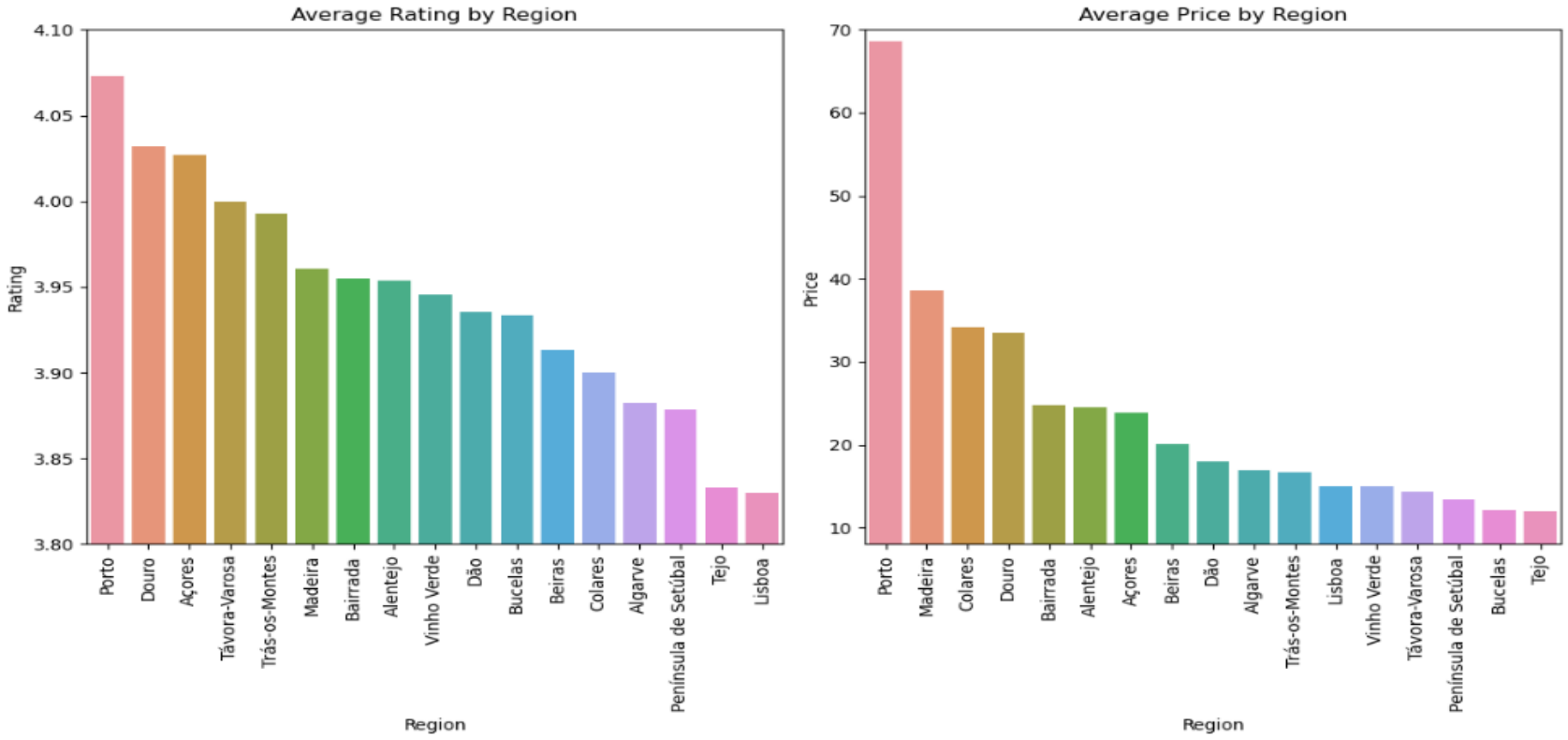
DATA ANALYSIS AND CONCLUSIONS

- The mean of the variable Rating was also calculated for each price category, which allows us to conclude that expensive wines tend to have better ratings:
    - Price < 5€: 3.64
    - 5€ <= Price <= 10€: 3.78
    - 11€ <= Price <= 15€: 3.92
    - 16€ <= Price <= 25€: 4.03
    - 26€ <= Price <= 50€: 4.11
    - Price > 50€: 4.36
- The best and worst rated wines in each price category were calculated. Overall, the worst rated wine was Vieira de Sousa Extra Dry White Port N.V. (2.7 average rating), and the best wines (4.8 average rating) were all in the last price category (> 50€):

	Winery	Year	Wine ID	Wine	Type	Rating	Num_Reviews	Price	Country	Region
1246	Krohn	1960	1136061	Colheita Port 1960	Fortified	4.8	62	223.58	Portugal	Porto
577	Luis Pato	2015	22317	Baga Beiras Quinta do Ribeirinho Pé Franco 2015	Red	4.8	105	159.00	Portugal	Beiras
2383	Ferreira	2018	75996	Vintage Port 2018	Fortified	4.8	187	88.15	Portugal	Porto
174	Douro Boys	2011	1639135	Cuvée 2011	Red	4.8	123	888.00	Portugal	Douro
1259	Taylor's	1966	1598998	Very Old Single Harvest Port 1966	Fortified	4.8	389	325.00	Portugal	Porto

DATA ANALYSIS AND CONCLUSIONS

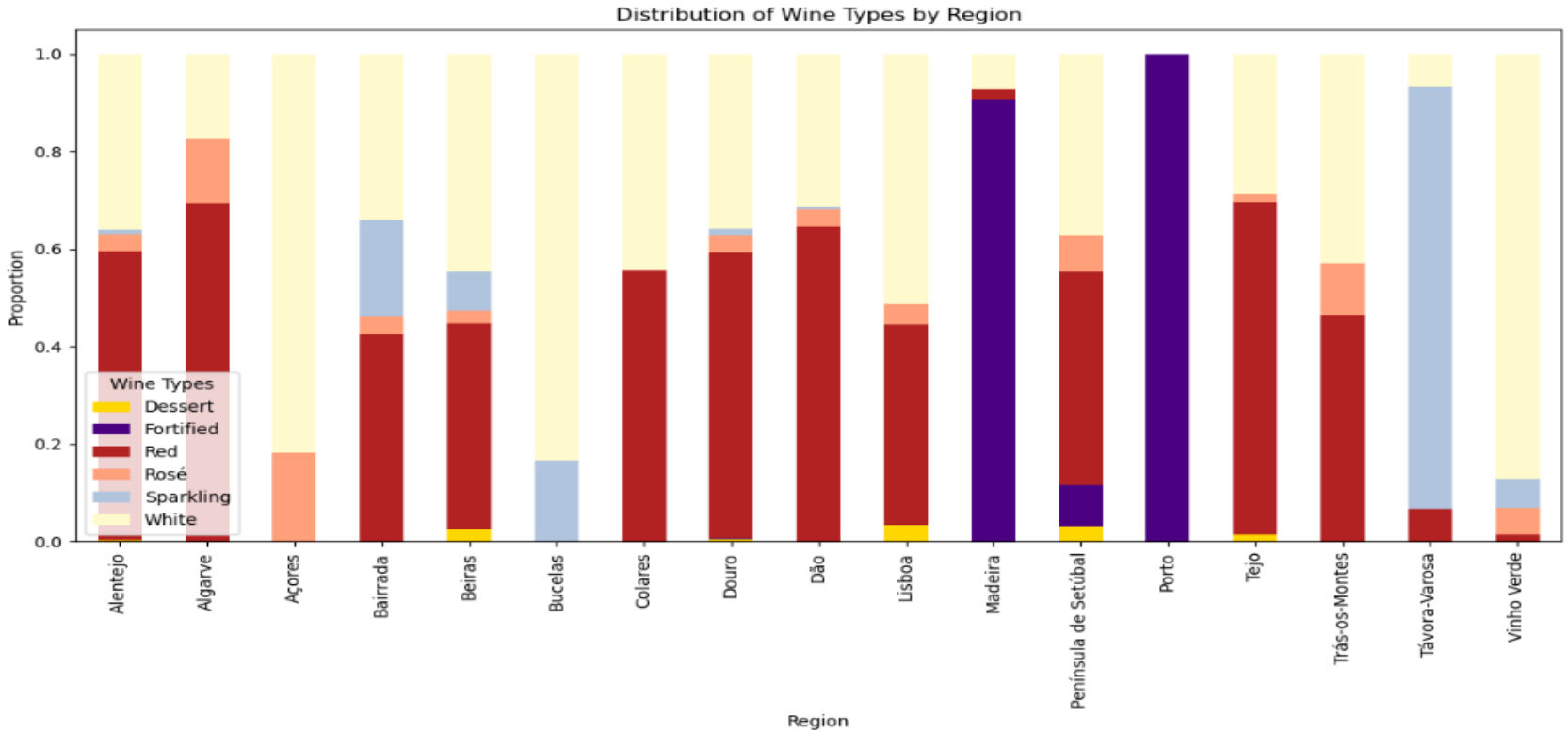
- Wine regions were compared and sorted by average rating and price: Porto, Douro, Açores and Távora-Varosa displayed an average rating >= 4.00 and Algarve, Península de Setúbal, Tejo and Lisboa displayed the lowest average ratings (<= 3.90). Porto, Madeira, Colares and Douro had the most expensive wines (>= 30€) and Tavóra-Varosa, Península de Setúbal, Bucelas and Tejo the cheapest wines (<= 15€):



- Using a normalized rating and normalized price, Porto was considered the best value region and Lisboa the worst.

DATA ANALYSIS AND CONCLUSIONS

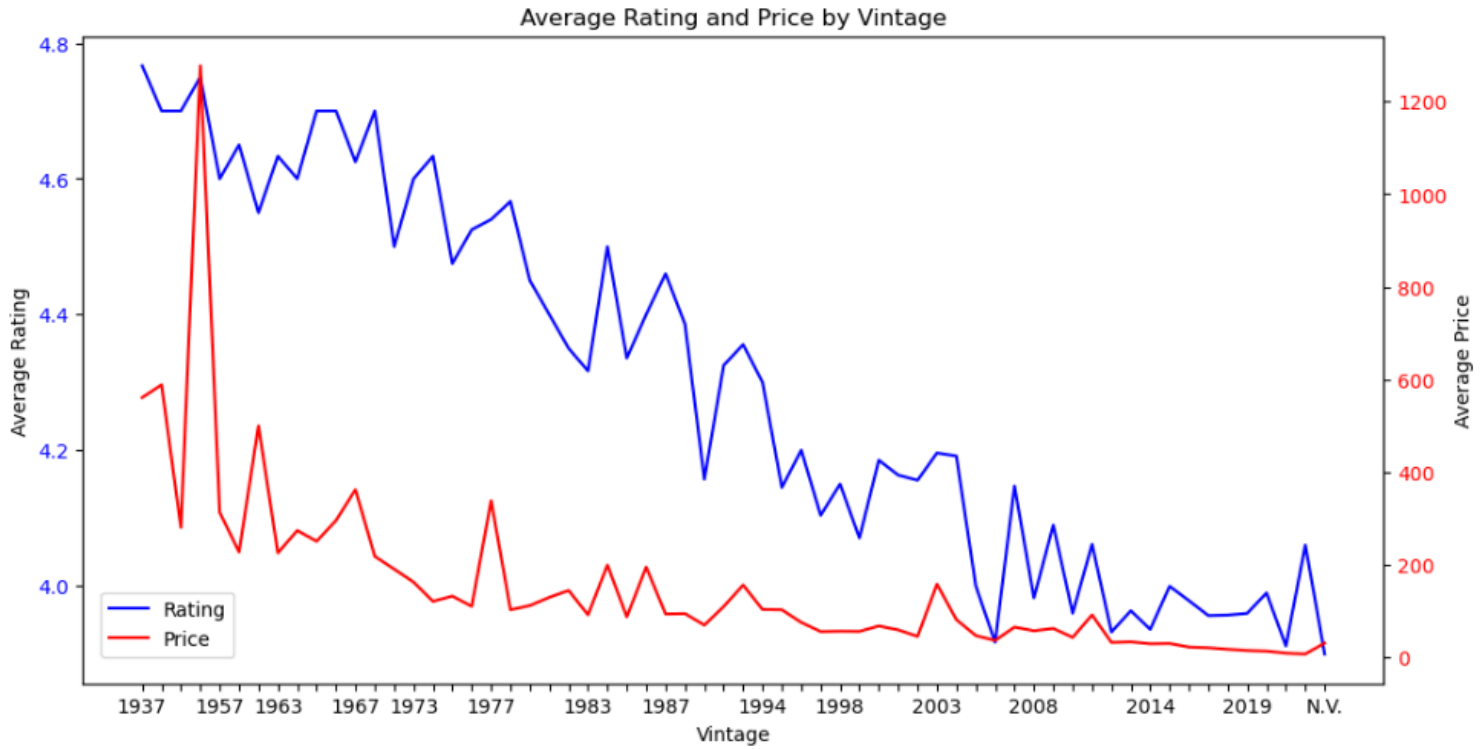
- The distribution of wine types by region was calculated:



- The average wine rating and price based on the type of wine was also calculated: Dessert wines were the top rated wines, followed by fortified and red wines. The most expensive wines were fortified, followed by red and dessert. However, there were only 17 dessert wines in the total, so that could explain the result. Rosé wines were the worst rated wines and the least expensive.

DATA ANALYSIS AND CONCLUSIONS

- The best and worst rated wineries (with at least 5 wine references) were calculated: the top three were Taylor's, Quinta do Vesúvio and Prats & Symington (their wines are from Douro and/or Porto), and the bottom three were Quinta de São Sebastião, Grão-Vasco and Casa de Santa Vitória.
- The average wine rating across different vintages (years) was calculated:



- The best vintages (with at least 50 ratings) were the following: 2007, 2011 and 2015.

DATA CLASSIFICATION AND CONCLUSIONS

- The wine rating (float) was converted to a category: less or equal than 3.6 – bad, between 3.7 and 3.9 – mediocre, between 4.0 and 4.2 – good, more than 4.3 – great.
- Models were trained and tested to predict the classification based on the following information: Winery/Producer, Year/Vintage, Type, Price, Number of Reviews, Region.
- Different algorithms were tested such as Decision Tree Classifier, Random Forest Classifier, Kneighbors Classifier, with different values for max leaf nodes, n estimators and n neighbors.
- The train/test size was split 70%/30%, with the following distribution:
    - BAD: 11.10% / 11.87%
    - MEDIOCRE: 33.66% / 33.50%
    - GOOD: 36.84% / 34.41%
    - GREAT: 18.37% / 20.20%
- As expected, “Price” proved to be the most important class for the models.
- The best performing model during training was the Random Forest Classifier, n_estimators = 48, “X” set to (“Winery”, “Year”, “Type”, “Price”, “Number of Reviews” and “Region”), which achieved 100%. However, this was a clear case of overfitting, as the same model only achieved approximately 59% during testing.
- The best performing model during testing was Decision Tree Classifier, max_leaf nodes = 24, “X” set to (“Winery”, “Year”, “Type”, “Price”, “Number of Reviews” and “Region”), which achieved approximately 61%. The same model with max_leaf nodes = 48 achieved better results during training (66% vs. 63%) but performed worse during testing (58% vs. 61%).

FINAL REMARKS

- Drawing on the insights gathered from this project, I would propose several enhancements to refine the data analysis and classification processes. Firstly, detailed information about the grape composition of each wine should be gathered. This information could include whether the wine is a blend or a varietal, and if the latter, the specific type of grape involved. Secondly, to achieve a more balanced and comprehensive representation, the datasets could be expanded by incorporating a wider variety of dessert wines, and wines from regions currently underrepresented. Lastly, a focused analysis of specific terminology in wine names, such as "Reserva", "Garrafeira", or "Vintage", could provide insights into any potential correlation between certain terms and a higher likelihood of receiving a better rating.

- In the study entitled, "Of Wines and Reviews: Measuring and Modeling the Vivino Wine Social Network," a pivotal conclusion was drawn: “The ratings and the reviews supplied by Vivino users display the same rich knowledge of wines as professional wine reviews. However, unlike the latter, Vivino users’ratings do not seem to be heavily affected by wine prices”. Contrarily, in the specific case of this analysis, which incorporated a constrained dataset of 3281 Portuguese wine references, the data analysis and the resulting classification models indicated a strong correlation between price and rating. For Portuguese wines, price could indeed be a determinant of the perceived quality, challenging the broader Vivino trend. Thus, while user reviews may remain largely uninfluenced by wine prices on a global scale, the relationship between cost and rating can vary substantially when the scope is narrowed down to a specific geographical context, as exemplified by Portuguese wines in this instance.