Predicting Ratings For a Variety of Dark Chocolates

University of British Columbia

Manvir Kohli, Julie Song, Kelvin Wong

2022-11-26

Contents

Summary

Four regression models were built using decision tree, support vector machine, ridge (linear regression) and k-nearest neighbors algorithms, and compared to determine the best model for predicting a chocolate's rating based on characteristics such as the number and type of ingredients in the chocolate, its amount of cocoa, location for the manufacturing company, memorable characteristics, and several others. The ratings are predicted based on a scale of 1 to 5. Our models performed well on a test set containing 759 observations, with most models having MAPE values of less than 10%.

Introduction

Commonly known as the "Food of the Gods", chocolate is a treat that is a world-wide favorite among the people. A recent study on food consumption changes during the COVID-19 pandemic and subsequent lockdown periods in Denmark, Slovania, and Germany revealed an increase in people consuming frozen and canned products, but also an increase in sweets, including chocolate (Janssen et al. (2021)). In particular, dark chocolate has been gaining attention due to the many health benefits that it provides (Montagna et al. (2019), Tan et al. (2021)).

There is interest in exploring this growing market of dark chocolate production, but there are a number of factors to consider about the chocolate itself. For example, dark chocolate is characterized by its large percentage of cocoa solids, which can vary between 50%-90% and gives it its signature bitter taste ("Dark Chocolate" (2022)). The project presented here is interested in predicting the rating for a type of dark chocolate on a scale of 1 to 5, given some of its characteristics, including its ingredients, amount of cocoa solids, manufacturing location, and more. This information can be useful for product analysis, and may be able to aid in predicting the popularity of a new product.

Our target mean absolute percent error (MAPE) is 5% for the predicted ratings. We chose MAPE because our rating scale is not very large and does not have units, so a relative percentage error is easier to understand

for this context.

Methods

Data

The data set is provided by the Manhattan Chocolate Society (@cocoa), and was found and retrieved from the tidytuesday data project (@tidytuesday), specifically through this link: https://github.com/rfordatascience/tidytuesday/blob/master/data/2022/2022-01-18/readme.md

The data set contains observations for different types of dark chocolate, including the manufacturing company and its location, the origin of the cocoa beans used to make the chocolate, the other ingredients in the chocolate, the amount of cocoa in the chocolate, and others. They have also provided a feature column that contains descriptive words relating to the characteristics of the chocolate flavor, and a final rating.

Before the models were built, the data also had to be processed. The information in the `ingredients` feature column was split apart into several columns, including one column for the number of ingredients in the chocolate, and one column for each of the recorded ingredients. The company location and cocoa bean origin features were modified such that only some categories that had total observations meeting a minimum threshold were considered, while the remaining categories were considered "Other". The feature for the amount of cocoa in the chocolate was also converted from a character datatype to a numerical datatype.

Analysis

Four regression models were built after splitting the original data set into training and testing data sets in a 70%-30% split. Apart from the processing performed above, some features were also dropped due to being practically unique to each observation, such as the reference number and the company manufacturer. Table 1 shows the hyperparameters that were considered for each regression model. Hyperparameter optimization was performed for each model, using 20 iterations of 5-fold cross-validation each.

Table 1: Model Hyperparameters

Model	Hyperparameter
KNN	Max Text Features
	Leaf Size
	Number of Neighbors
	Weights
Ridge	Max Text Features
	Alpha
SVM RBF	Max Text Features
	C
	Gamma
Decision Tree	Max Text Features
	Max Depth
Random Forest	Max Text Features
	Max Depth
	Number of Estimators

The initial exploratory data analysis for this dataset was performed in R (R Core Team (2022)) using the following packages: dplyr (Wickham et al. (2022)), tidyverse (Wickham et al. (2019)), docopt (de Jonge (2020)), cowplot (Wilke (2020)), knitr (Xie (2014)), and kableExtra (Zhu (2021)).

The modelling was performed in Python (Van Rossum and Drake (2009)) using the following packages, available in the provided environment file: matplotlib (Hunter (2007)), scikit-learn (Pedregosa et al. (2011)), pandas (McKinney (2010)), imbalanced-learn (Lemaître, Nogueira, and Aridas (2017)), lightgbm (Ke et al. (2017)), joblib (Joblib Development Team (2020)), docopt (Keleshev (2014)), and dill (Van Rossum (2020)).

Results and Discussions

The distributions for the numerical features and their pairwise correlations are shown in Figures 1 and 2, which reveals the that there are no strong relationship between any of our numerical features. Figure 3 shows the distributions for the categorical features. The top company locations which had at least 20 observations and the top bean origin countries with at least 10 observations were kept, while the remaining categories in each feature were grouped as Other. It is revealed that the manufacturing company feature has an overwhelming number of categories, and even taking into considering only the top 50 companies, the Other category has far more points compared to the others. This feature was considered too unique to the observations, and was omitted from the modelling. The boxplots in Figure 4 highlight some of the categories that have tendencies towards higher ratings, such as the presence of different ingredients or the company location. For example, chocolates produced in Austria and Denmark appear to have higher ratings on average compared to the rest of the group.

Figure 1: Numeric Plots

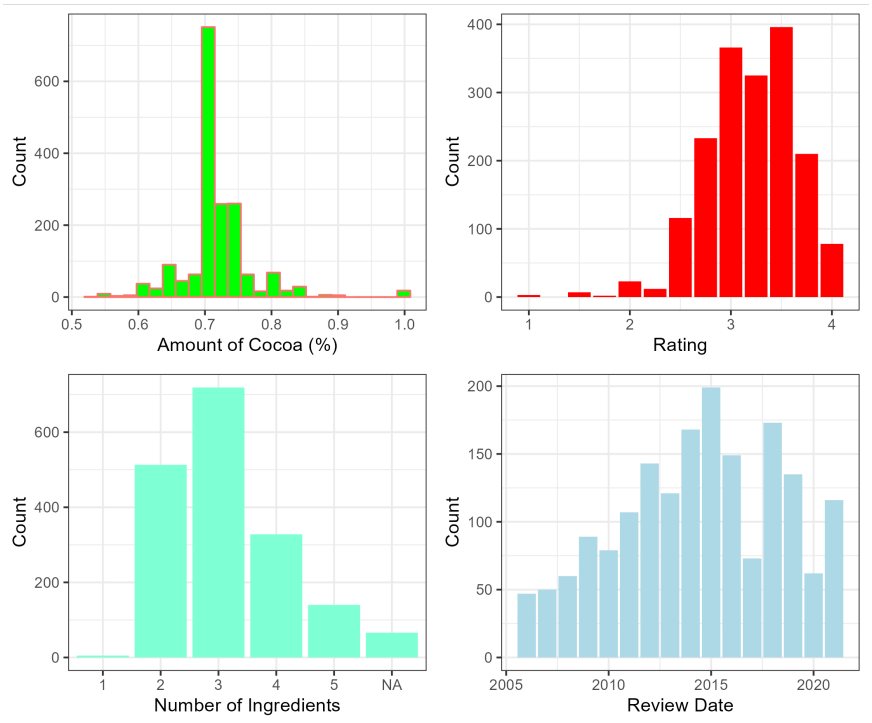

Figure 1: Numerical feature distributions

Figure 2: Pairwise Scatter Plots

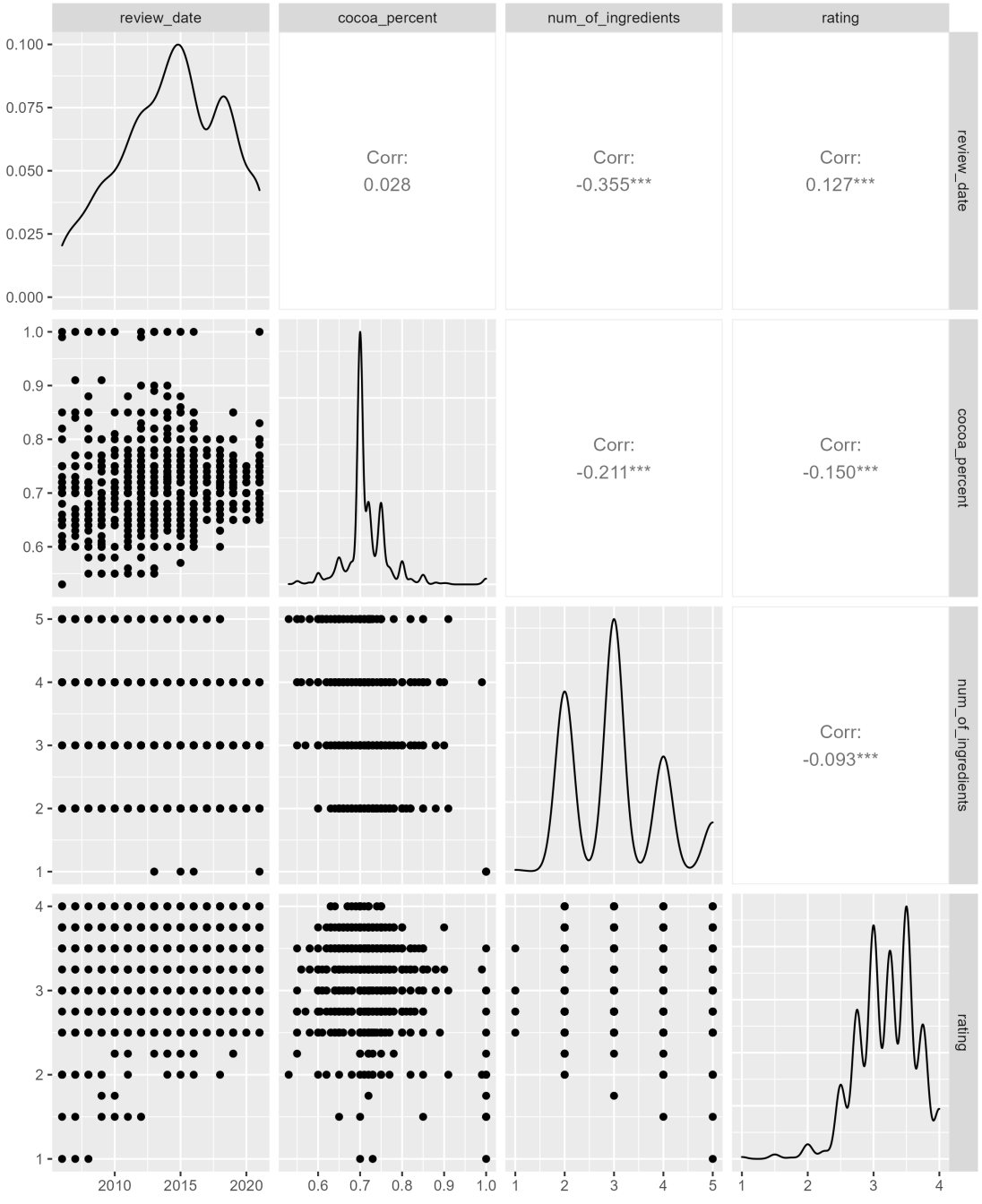

Figure 2: Numerical feature scatterplots

Figure 2: Categorical Plots

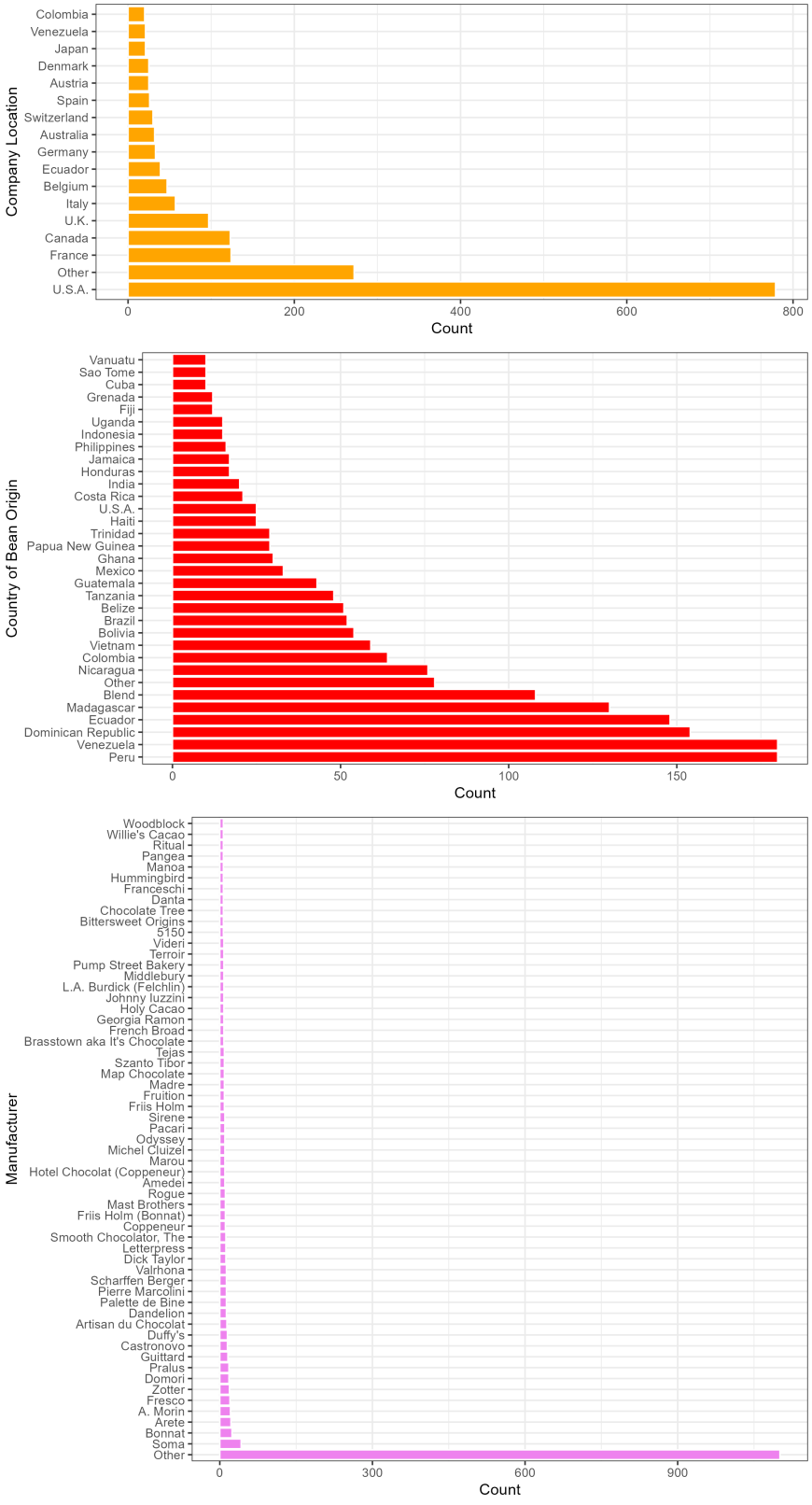

Figure 3: Categorical feature distributions

Figure 4: Boxplots by Rating

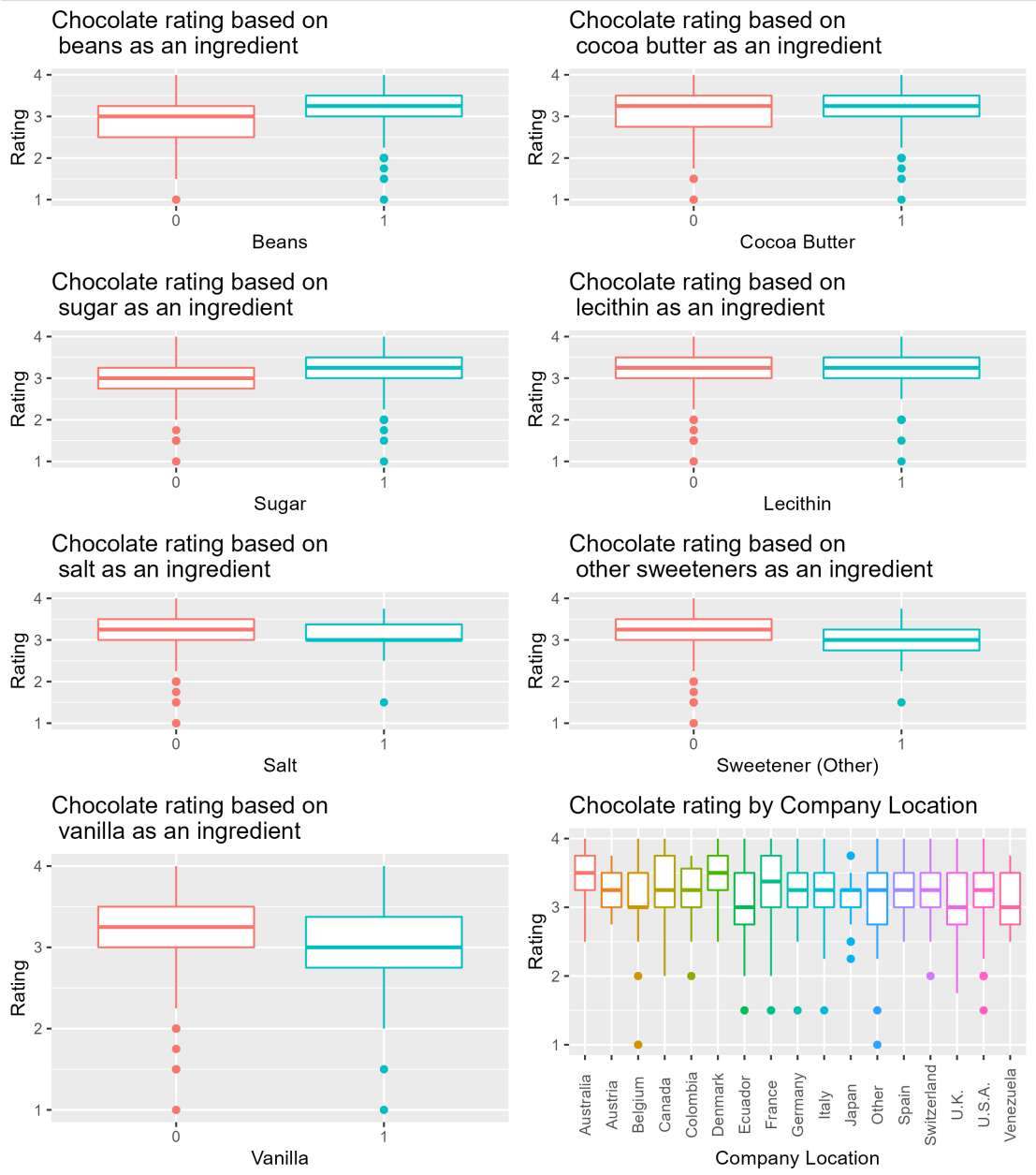

Figure 4: Categorical feature boxplots

Table 2 shows the optimal hyperparameters, their associated values, and the resulting validation scores for each model. The scoring metric was chosen to be based on the mean absolute percentage error. It can be seen that the KNN model may be overfitting, since its training score is 0 (i.e. a perfect model), while its validation score indicates 10.6% error. The decision tree model also had higher validation error compared to the other models, perhaps because decision tree models can be sensitive to feature values, such that slight changes in the values can result in very different results. It is also worth noting that although the random forest model performed better than these two models, it required a much longer amount of time to run,

which is quite undesirable for practical usage of the model.

Table 2: Optimized Hyperparameters With MAPE Values

Model	Hyperparameter	Optimized Value	Mean Validation MAPE (%)	Mean Training MAPE (%)
KNN	Max Text Features	395	11.0	0.0
	Leaf Size	450		
	Number of Neighbors	13		
	Weights	Distance		
Ridge	Max Text Features	437	9.1	6.7
	Alpha	1		
SVM RBF	Max Text Features	262	9.0	6.8
	C	39.5		
	Gamma	0.0014		
Decision Tree	Max Text Features	176	10.9	8.6
	Max Depth	9		
Random Forest	Max Text Features	396	9.5	3.4
	Max Depth	78		
	Number of Estimators	351		

The models performed well on the test data sets, with similar results compared to the validation scores, as shown in Table 3. The best-performing model was the SVM RBF model, as it only has an average of 8.6% error for its ratings, while the worst-performing model was the decision tree model, with an average of 11% error. Our best model was close to our target MAPE of 5%, although its error is still a bit high. Although this is not extremely different, since our rating scale is quite small, we would still prefer to decrease the error if possible.

Table 3: Test Results for Each Model

Model	Test MAPE (%)
KNN	10.6
Ridge	8.7
SVM RBF	8.6
Decision Tree	11.1
Random Forest	9.2

There are several areas for improvement with these models. For instance, perhaps we could do some feature engineering by using a sentiment analysis package on the memorable characteristics text feature, and manually aggregate some of the similar terminology (such as “acidic” and “acid”). Since most of our features are categorical features, we could also do some feature crosses and determine if there are any relations between the features. This might require some more careful examination of the data, but for instance, we could begin with some crosses of the ingredients present in the chocolate. As well, we could introduce a noise attribute for regularization, as it may help us identify exactly how well the model is actually performing with these features. Finally, we may also do some feature selection based on coefficient weightings, for instance for the Ridge model. Perhaps these can allow us to make the models more generalizable, and better able to predict well for brand-new types of chocolate, which is the ultimate goal for the project.

References

"Dark Chocolate." 2022. *The Nutrition Source*. Harvard T.H. Chan School of Public Health. https://www.hsph.harvard.edu/nutritionsource/food-features/dark-chocolate/.

de Jonge, Edwin. 2020. *Docopt: Command-Line Interface Specification Language*. https://CRAN.R-project.org/package=docopt.

Hunter, J. D. 2007. "Matplotlib: A 2d Graphics Environment." *Computing in Science & Engineering* 9 (3): 90–95. https://doi.org/10.1109/MCSE.2007.55.

Janssen, Meike, Betty P. I. Chang, Hristo Hristov, Igor Pravst, Adriano Profeta, and Jeremy Millard. 2021. "Changes in Food Consumption During the COVID-19 Pandemic: Analysis of Consumer Survey Data from the First Lockdown Period in Denmark, Germany, and Slovenia." *Frontiers in Nutrition* 8. https://doi.org/10.3389/fnut.2021.635859.

Joblib Development Team. 2020. *Joblib: Running Python Functions as Pipeline Jobs*. https://joblib.readthedocs.io/.

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. "Lightgbm: A Highly Efficient Gradient Boosting Decision Tree." *Advances in Neural Information Processing Systems* 30: 3146–54.

Keleshev, Vladimir. 2014. *Docopt: Command-Line Interface Description Language*. https://github.com/docopt/docopt.

Lemaître, Guillaume, Fernando Nogueira, and Christos K. Aridas. 2017. "Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning." *Journal of Machine Learning Research* 18 (17): 1–5. http://jmlr.org/papers/v18/16-365.html.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python." In *Proceedings of the 9th Python in Science Conference*, edited by Stéfan van der Walt and Jarrod Millman, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a .

Montagna, Maria Teresa, Giusy Diella, Francesco Triggiano, Giusy Rita Caponio, Osvalda De Giglio, Giuseppina Caggiano, Agostino Di Ciaula, and Piero Portincasa. 2019. "Chocolate, 'Food of the Gods': History, Science, and Human Health." *International Journal of Environmental Research and Public Health* 16 (24). https://www.mdpi.com/1660-4601/16/24/4960.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–30.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Tan, Terence Yew Chin, Xin Yi Lim, Julie Hsiao Hui Yeo, Shaun Wen Huey Lee, and Nai Ming Lai. 2021. "The Health Effects of Chocolate and Cocoa: A Systematic Review." *Nutrients* 13 (9). https://www.mdpi.com/2072-6643/13/9/2909.

Van Rossum, Guido. 2020. *The Python Library Reference, Release 3.8.2*. Python Software Foundation.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wilke, Claus O. 2020. *Cowplot: Streamlined Plot Theme and Plot Annotations for 'Ggplot2'*. https://CRAN.R-project.org/package=cowplot.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.