DIMITRIOS POULIMENOS

200291237

Data Science

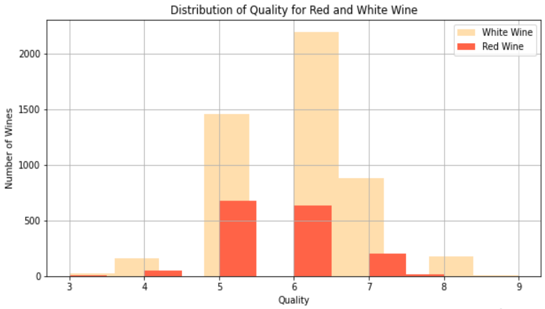
Newcastle University

Abstract

This report investigates how the quality of the different varieties of Vinho Verde wine can be predicted and what determines this. In order to gain a better understanding, the given datasets were visualised. In addition, the confusion matrices were generated to identify the most important predictor variables of producing the best quality of wine and thus three different models using logistic regression, decision tree and random forest methods have been created. The analysis of these methods revealed that strong correlation between quality and alcohol along with sulphates in both wines. Additionally, the accuracy scores of the analysis showed that the random forest method is the best model to predict quality. In conclusion, red wine quality can be affected by alcohol, sulphates, citric acid, and fixed acidity, unlike white wine whose quality can be affected mainly by alcohol and the best method to predict that is random forest for both wines.

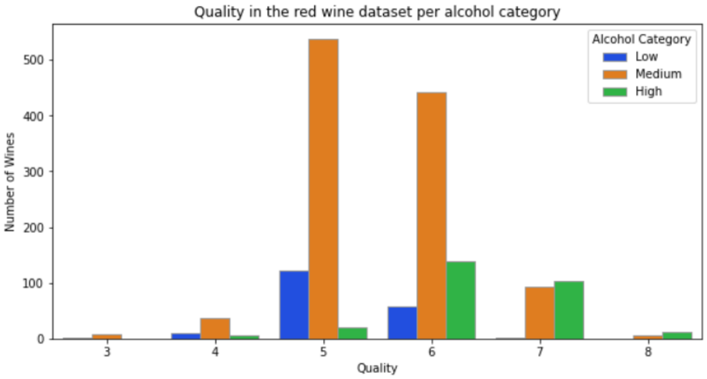
What Was Done and How

Exploring each of the datasets is an essential part of data science to gain a better understanding of them. Initially, data aggregation had to be done by using a python data analysis library called pandas (McKinney, 2008). To get the most out of datasets, two data frames have been created for red and white wine respectively for further usage.

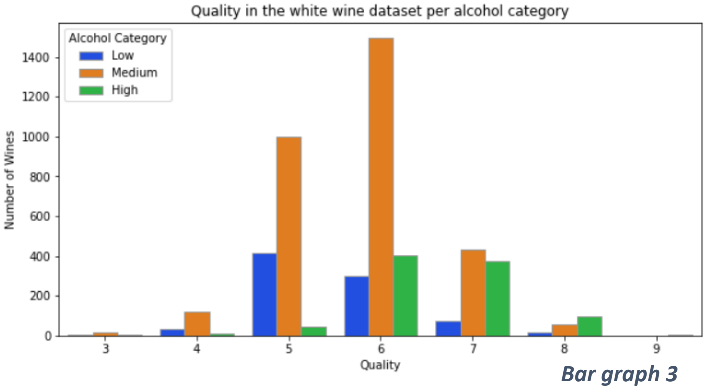


Bar graph 1

After reading, aggregation and data frame creation three plots had to be created for the demonstration of quality distribution for both datasets. This was done by using a python statistical data visualization library called seaborn (Waskom, 2021) and another library which called matplotlib (Droettboom, 2003). After demonstrating the quality distribution of each dataset then both datasets had to be visualised on the same histogram (Bar graph 1). Afterwards, to discretise the alcohol content of both datasets, a new predictor has been created named alcohol_cat with which the wine can be categorized into three different categories which are low, medium, and high. This had to be done to illustrate the quality of wines in different categories of alcohol (Bar graph 2 & 3, page 1). Then it was necessary to create a new column which contains whether a wine is dry or sweet. This was achieved using the residual sugar predictor and thus the isSweet column created containing the wines was classified into dry and sweet according to the percentage of residual sugar contained in each wine. Finally, by visualizing the number of dry and sweet wines, it will be easy to use this element later to understand the percentage of wines that have the highest quality between dry and sweet wines (Bar graph 4 & 5, page 2).

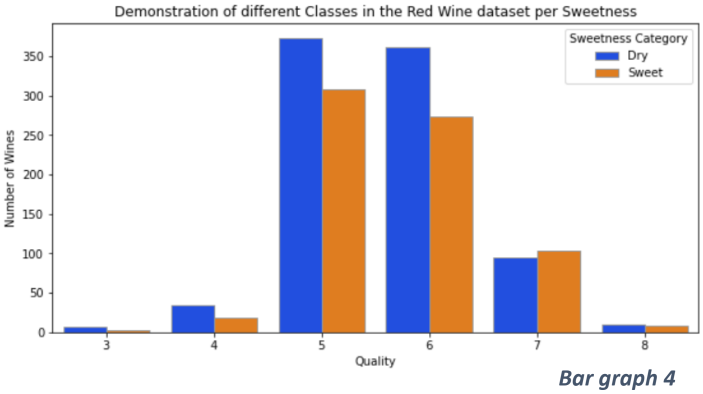


Bar graph 2

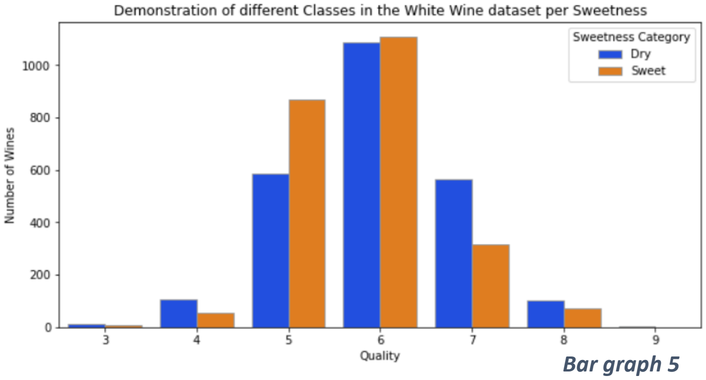


Bar graph 3

After exploring and plotting the datasets, it was time to create a correlation heatmap so that it can be understood what the relationship of each predictor with the rest is but specifically with the quality. This was done in order to highlight the predictors that are responsible for the quality of each wine depending on its percentage (Heatmap 1 & 2, page 2 & 3).

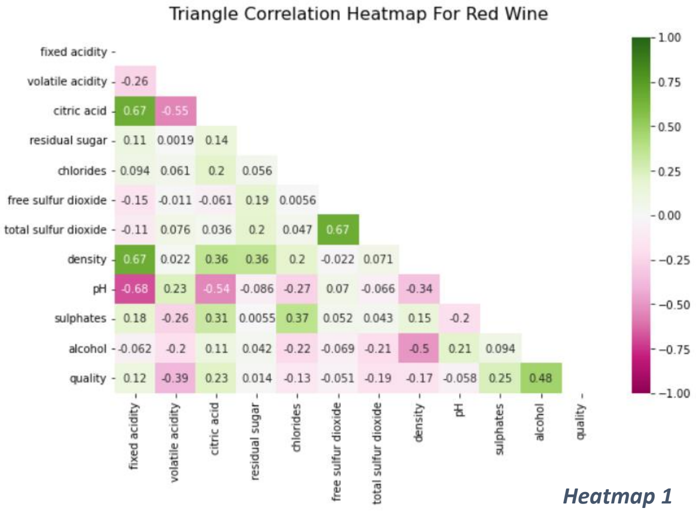


Bar graph 4



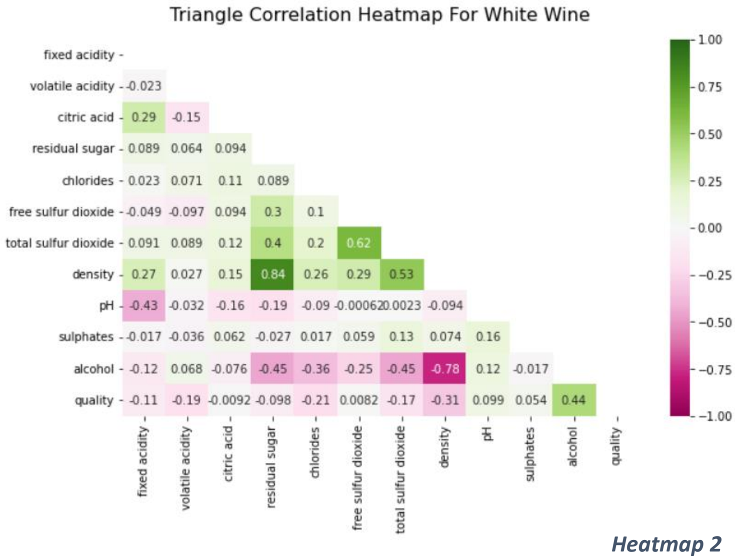
Bar graph 5

Next, the most crucial part of the assignment was to prepare the datasets in order to create a model to predict the quality of the given wine. Considering this as a classification problem a new column called quality_cat had to be created with which it would be possible to separate the wines into two types of quality, the low and the high. By setting the lowest quality to zero and the highest quality to the highest recorded quality and using the pandas library it was possible to create this new column. Then after the three different columns were created to categorize the quality, alcohol and dryness of the wine, it was the turn of converting these columns to binary classification in order to use them for the methods that will follow. So, the category alcohol_cat was changed from low, medium, and high to 0, 1 and 2 respectively, while the categories quality_cat and isSweet were replaced by two new categories for each of them named low quality, high quality and dry, sweet respectively (Data frame 1, page 3).



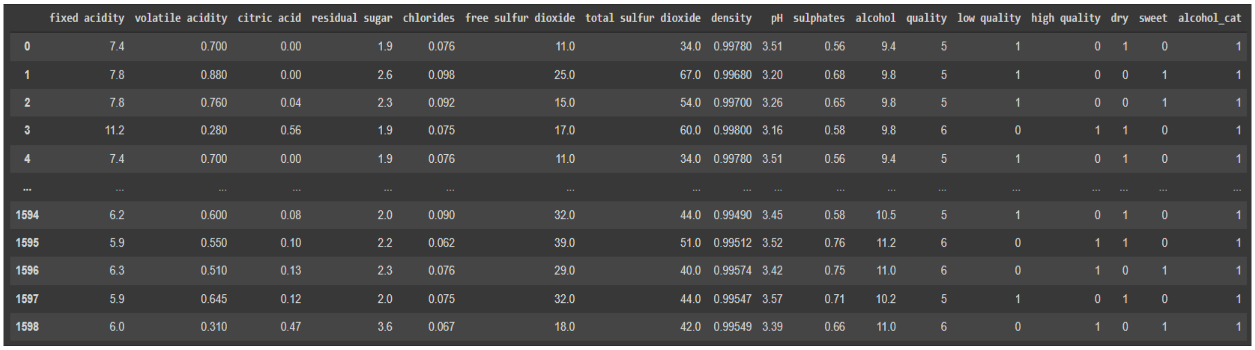
Heatmap 1

Then the dataset was divided into a training set and a test set so that the model could be trained. Most of the data has been used for training, and a smaller portion of the data has been used for testing. This part is important for evaluating data. In addition, the datasets have been balanced in order to generate higher accuracy models, higher balanced accuracy along with balanced detection rate. Thus, is important to have a balanced dataset for a classification model in general. Next, data scaling took place to make it easy for the model to learn and understand the problem. After the preparation of the models, it was the turn of the actual methods to be implemented. Three different methods have been used to choose the more efficient and accurate of them. The first one was the logistic regression method which is easier to implement and efficient enough to train the model. Also, model coefficients can be interpreted as an indicator of the importance of features. The second method that has been used was the decision tree which in general is easy to interpret for small-sized datasets and is inexpensive to construct in terms of time. The last method that has been used was the random forest method which can handle the dataset containing categorical variables in the case of classification and overall, it performs better results for classification problems.



Heatmap 2

After the implementation of the models, evaluation took place. Evaluation of each model was made possible by using the same format for all three models. At first, the logistic regression method has been evaluated. With a classification report and a confusion matrix as well and then with a ROC diagram. Unfortunately, the implementation of cross-validation wasn't possible for this method due to a lack of capabilities. Next, the evaluation of the decision tree method took place and like the previous method a classification report, confusion matrix, and ROC diagram were held. In addition, in this model, cross-validation cannot be implemented. And finally, in the random forest method where all types of evaluation were successfully performed. The purpose of the evaluation was to quantify the quality of a model's predictions.



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality	low quality	high quality	dry	sweet	alcohol_cat
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5	1	0	1	0	1
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5	1	0	0	1	1
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5	1	0	0	1	1
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6	0	1	1	0	1
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5	1	0	1	0	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1594	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5	1	0	1	0	1
1595	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76	11.2	6	0	1	1	0	1
1596	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6	0	1	0	1	1
1597	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5	1	0	1	0	1
1598	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6	0	1	0	1	1

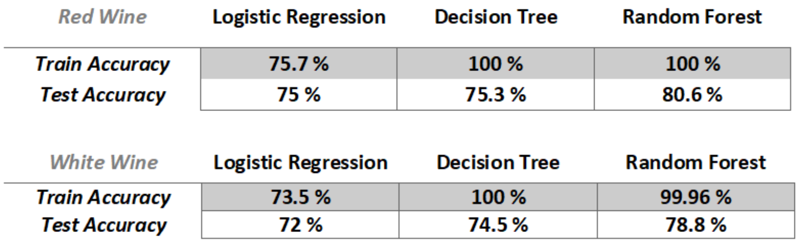
Data frame 1

Results and Evaluation

Here will be an analysis of the results of the artefacts along with the data produced by the three models to understand more of these three models based on their analysis and which of these three models best serves the specific scenario. Starting the visualization of datasets, it is obvious that in red wine the largest number of wines is in the low-quality category in contrast to white wine where most wines are in the high-quality category. This shows us that if from now on we had to choose based on the quality we would immediately choose white wine but there are many parameters which will be analyzed later, making this choice a little more complicated. Going forward, what it can be seen is that the higher the quality of the wine, the higher the percentage of alcohol. The same seems to happen with white wine, showing us that the quality is directly affected by the alcohol content of the wine. Then, dividing the wines according to the amount of residual sugar they contain, it is obvious that in the red wine from the **1597** samples, **881** are dry, making the dry wines predominant. Like red wines and white wines were from **4896** to **2467** are dry. Thus, comparing the dryness of the wines with the quality, it is obvious that the higher the quality of the red wines, the less the dry wines, making the sweets superior in quality in contrast to the white wines, where the number of dry wines decreases as the quality but in the end dry wine is what continues to dominate.

Furthermore, after the creation of the correlation heatmaps, the relation of all the components of the wine with the quality can be seen. The prices produced to show that in red wine the biggest correlation has the predictors alcohol (0.48), **sulphates (0.25)**, **citric acid (0.23)**, **fixed acidity (0.12)**, and **residual sugar (0.014)** in contrast to white wines where the predictors that can affect quality are clearly fewer and with a lower number of positive correlations. speaking of which, **alcohol (0.44)**, **ph (0.099)**, **sulphates (0.054)**, and **free sulfur dioxide (0.0082)**. The common predictors in both wines are alcohol and sulphates. Thus, by trying different prices of the above in both categories of wines can be produced superior quality.

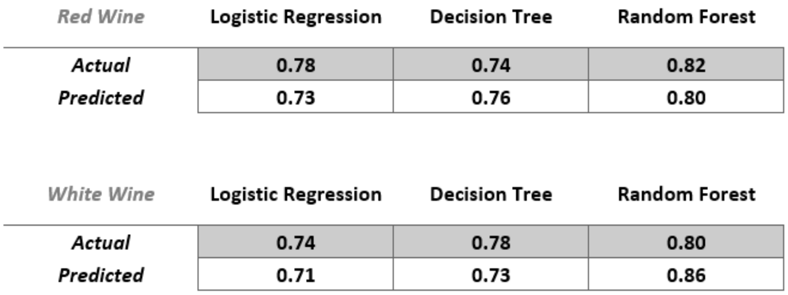
In addition, the results of the created models should be examined so that it can decide which of the three models is the most suitable. Initially, for the red wine the logistic regression (Lawton, 2022) method, as it can be seen from the classification report the f1-score is at **0.75**, a number which is the same as the result of the classification report of the decision tree (**0.75**) but less than the classification number report of random forest (**0.81**). On the other hand, the white wine's classification report number for logistic regression method is (**0.72**), for decision tree (**0.75**), and for random forest (**0.79**) (Table 1, page 4). Overall, the classification report (Kharwal, 2021) (Krishnan, 2018) serves to highlight the model with the highest prediction quality giving us the number of true positives, true negatives, false positives, and false negatives. It is the turn of the confusion matrix for the red wine, from which it can be observed that for the logistic regression, the percentage of successful forecasts is quite close to the real values (**actual-0.78, predicted-0.73**). Similar, the numbers of successful predictions for the decision tree method are even closer to the real ones having a deviation of 0.02 (**actual-0.74, predicted-0.76**) the deviation is the same for the random forest method but with larger numbers (**actual-0.82, predicted-0.80**). For the white wine, the number from the confusion matrix for logistic regression is (**actual-0.74, predicted-0.71**), for decision tree (**actual-0.78, predicted-0.73**) and lastly for random forest (**actual-0.80, predicted-0.76**) (Table 2, page 4). The usefulness of the confusion matrix (aniruddha, 2020) lies in measuring efficiency and providing a better picture of how effective each model is or not.



Red Wine	**Logistic Regression**	**Decision Tree**	**Random Forest**
Train Accuracy	**75.7 %**	**100 %**	**100 %**
Test Accuracy	**75 %**	**75.3 %**	**80.6 %**

White Wine	**Logistic Regression**	**Decision Tree**	**Random Forest**
Train Accuracy	**73.5 %**	**100 %**	**99.96 %**
Test Accuracy	**72 %**	**74.5 %**	**78.8 %**

Table 1



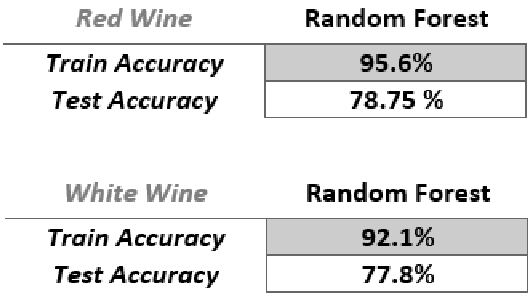
Red Wine	**Logistic Regression**	**Decision Tree**	**Random Forest**
Actual	**0.78**	**0.74**	**0.82**
Predicted	**0.73**	**0.76**	**0.80**

White Wine	**Logistic Regression**	**Decision Tree**	**Random Forest**
Actual	**0.74**	**0.78**	**0.80**
Predicted	**0.71**	**0.73**	**0.86**

Table 2

Next is the ROC-AUC curve (Narkhede, 2018) for each model which is another way in which it can be understood whether a model is capable of distinguishing between classes. The roc curve represents the possibility, and the AUC curve represents the measure of separability. So, the higher the AUC the better the model in predictions. For red wine the ROC diagram of the logistic regression method is **0.82** in which the base of the diagram can be seen to have an upward trend, for the decision tree method it is at **0.75** while for the random forest method at **0.90** with an equally upward trend. As far as white wine is concerned, it is obvious that the numbers are smaller than red wine, except in the case of the decision tree method in which it remained the same (**0.75**). Regarding the logistic regression method, it is at **0.80** and for the random forest method at 0.89.

Finally, there is cross-validation (Brownlee, 2018) (Goyal, 2021) which shows the performance of the model. It also prevents the model from overfitting, which makes it even more useful for the analysis of each model. Unfortunately for the logistic regression method, I could not implement cross-validation. For the decision tree method, it is not possible to implement cross-validation as there are no hyperparameters. And so, we come to the random forest method in which the results are showed in table 3 in this page. This numbers shows that the overfit that existed in now reduced and thus the number are the best for this model.



Red Wine	**Random Forest**
Train Accuracy	**95.6%**
Test Accuracy	**78.75 %**

White Wine	**Random Forest**
Train Accuracy	**92.1%**
Test Accuracy	**77.8%**

Table 3

Starting with the evaluation, in the first task, it was asked to aggregate and then create plots from the two wine datasets in order to demonstrate the quality distribution, which was implemented successfully. Then, the creation of new columns was requested, which illustrate the category of alcohol in each wine the quality category of each wine as well as the sweetness category of each wine. However, it would probably be more useful to use one column instead of two for the new categories. Next, a correlation matrix was asked to be created which was created and the correlations can be distinguished successfully which makes it successful. Then it was requested to prepare the data frames to create the prediction models. This part was successful as three prediction models were created, and the corresponding results were produced. Finally, the requested part of the analysis was also successful based on the results that have been produced. Overall, each part of the assignment implemented as requested except the cross-validation of the logistic regression method which was quite hard to understand and implement.

Conclusions

In conclusion, two datasets of white and red wines from the Vinho Verde variety were used to create a prediction model for the best quality. After the creation of new predictors and the emergence of predictors that play the most important role in the quality of a wine, the data were prepared for the creation of the three models which were created and then the analysis of these models took place in order to highlight what predicts more precisely the quality of each wine.

After my friction with this assignment, in future I would change some things from which I realized that they could be implemented in a different way and more efficiently. First, I would not create two subcategories for each new category I generated, namely quality_cat and isSweet. Secondly, I would approach the exercise as a regression problem and not as a classification because there are some very nice elements that I left unexplored, and I could have learned a lot from them. Thirdly, I would spend more time reading in detail about the topic, which I did not do and any shortcomings in the code and the report could have been avoided.

In my opinion, I applied what I was asked by this project except for extensions and cross-validation in the logistic regression method. The visualization of the data was my strong point while in the creation but also the analysis of the models I believe that I could have done better if I had understood the topic 100%. Overall, it was a very difficult project and at the same time a challenge for me which I was able to cope to the best possible degree.

References

aniruddha. (2020). *Analytics Vidhya*. Retrieved April 17, 2020, from https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/#:~:text=A%20Confusion%20matrix%20is%20an,by%20the%20machine%20learning%20model.

Brownlee, J. (2018). *Machine Learning Mastery*. Retrieved March 23, 2018, from https://machinelearningmastery.com/k-fold-cross-validation/#:~:text=Cross%2Dvalidation%20is%20primarily%20used,the%20training%20of%20the%20model.

Cournapeau, D. (2007). *scikit-learn*. Retrieved June 2007, from https://scikit-learn.org/

Droettboom, M. (2003). *Matplotlib*. Retrieved 2003, from https://matplotlib.org/

Goyal, C. (2021). *Analytics Vidhya*. Retrieved May 21, 2021, from https://www.analyticsvidhya.com/blog/2021/05/importance-of-cross-validation-are-evaluation-metrics-enough/#:~:text=%F0%9F%91%89%20k%2DFold%20Cross%2DValidation%3A&text=It%20ensures%20that%20the%20score,repeated%20k%20number%20of%20times.

Kharwal, A. (2021). *Thecleverprogrammer*. Retrieved July 17, 2021, from https://thecleverprogrammer.com/2021/07/07/classification-report-in-machine-learning/#:~:text=A%20classification%20report%20is%20a,of%20your%20trained%20classification%20model.

Krishnan, M. (2018). *Muthukrishnan*. Retrieved July 7, 2018, from https://muthu.co/understanding-the-classification-report-in-sklearn/#:~:text=A%20Classification%20report%20is%20used,classification%20report%20as%20shown%20below.

Lawton, G. (2022). *TechTarget*. Retrieved January 10, 2022, from https://www.techtarget.com/searchbusinessanalytics/definition/logistic-regression

McKinney, W. (2008). *pandas*. Retrieved January 11, 2008, from https://pandas.pydata.org/

Narkhede, S. (2018). *Medium*. Retrieved June 16, 2018, from https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Skyl.ai. (2019). *Medium*. Retrieved September 10, 2019, from https://medium.com/@skyl/evaluating-a-machine-learning-model-7cab1f597046#:~:text=Model%20Evaluation%20is%20the%20process,data%20with%20it's%20own%20predictions.

Waskom, M. (2021). *seaborn*. Retrieved from https://seaborn.pydata.org/