Wineinformatics: Determination of Score Category Using K-Nearest Neighbor with Various Distance Metrics

Nnaemeka Okereafor	Jonathan Onyumbe	Alex Ridley	Zeqing Dong	Bernard Chen
Computer Science University of Central Arkansas Conway, Arkansas, USA nokereafor1@cub.uca.edu	Computer Science University of Central Arkansas Conway, Arkansas, USA	Computer Science University of Central Arkansas Conway, Arkansas, USA	Computer Science University of Central Arkansas Conway, Arkansas, USA zdong1@cub.uca.edu	Computer Science University of Central Arkansas Conway, Arkansas, USA bchen@uca.edu

***Abstract*-** Wineinformatics is an emerging domain in the field of data science. Wine production is multivariate, in that the final quality of the wine is determined by a multitude of factors. One of the fundamental determinants for the quality of a given wine is its "terroir" -- the characteristics of the soil where the wine's ingredients are cultivated. Based on previous research, we know that the descriptive qualities for a given wine can be represented as a binary tuple, e.g. "bubbly = 1, high_tannin = 0, strong_finish = 1" through the Computational Wine Wheel. We have also determined that the score or grade which are either greater than 90 or less than 90 -- can be predicted by combining distance metrics with K-Nearest-Neighbor (KNN). This paper surveys the classification accuracy of a wine dataset for various distance metrics with KNN. These results are cross-validated using the leave-one-out method for different values of k.

Keywords*: *Wineinformatics, Bordeaux Wine, Computational Wine Wheel, Data Mining, KNN

1. Introduction

As it is defined, data mining is part of data science which involves having a domain knowledge to discover useful information from domain-related data. Therefore, the domain knowledge has high importance in the accomplishment of this research because first you need to have the domain problem, that will lead you to find out useful information in that domain. Our domain here is Wineinformatics which was introduced in the previous project [1], it is basically the understanding of wine to serve as the domain knowledge. Wineinformatics is a new field of research that needs more attention especially with all vast number of wine consumer or producer, there's a lot of things that can be accomplished in this field that will benefit the society.

There's this company well known for its evaluation of wines which is called "Wine Spectator", its focus is to evaluate wines by certain number of experts in that domain [4]. A total of about 400,000 wine reviews have been published by the company, and the magazines will be distributed in 15 issues per year, with between 400 and 1000 wine reviews per issue. The test expert used a blind test that only knows the year and type of the wine when they test it. The evaluation of the wine used by Wine Spectator has a scale of 50-100 score scale [2]. Therefore, with that scoring scale, it creates an influence or makes a big impact in a sense that it gives any wine consumer and most importantly the buyer a pretty good estimate value of that wine with a given score.

In this research, we decided to focus on the famous collection of Bordeaux wines. First, Bordeaux is considered as one of the most popular regions of wines. Therefore, it is considered at high standard compared to other wines from different regions. It is also well known for its vintage quality of wines. It is considered as one of the biggest regions of wine productions in the world [5]. It has a lot of attention and attract a vast number of consumers to that city. In our research, we focused on the Bordeaux wines reviews from year (2010 – 2016) which were extracted from Wine Spectator [2]. Then the reviews were transformed from human languages into a more readable computer related form using the Computation Wine Wheel [6]. For that matter, it leads to the point of understanding how those wines are being reviewed and by who.

The goal of this research is to determine either a wine belongs to the 90+ category or the 90- category. For that reason, we decided to use a white box data mining algorithm called K-Nearest Neighbor (k-NN) which is "a non-parametric method used for classification and regression" [3]. In both cases, the input consists of the closest examples of K training in the feature space. The algorithm's output is a class membership, and an object is categorized by its neighbors ' majority vote, assigning the object to the most common class among its nearest k neighbors (k>0). In other words, no model is built by k-NN. K values are selected, and the algorithm calculates instance distances and then directly predicts labels. As established in previous work [1,3], the *k-nearest-neighbor* (KNN) algorithm can be used in combination with various distance metrics to accurately determine the score category for those Bordeaux wines. Those wines can either be put in the 90+ or 90 – score category. For that reason, in our research we will evaluate those distance metrics and find the one that gives the best accuracy across all the K values that will be used.

2. Wine data

Wineinformatics is a subset of data mining involving the extraction of useful information from wine datasets. In previous work, the "Computational Wine Wheel" was developed to automatically process wine reviews and extract useful descriptive qualities [2]. This tool was used to process thousands of online reviews from the lifestyle magazine *Wine Spectator*, which has been publishing wine reviews since 1976. *Wine Spectator* provides a free and easily searchable online database of over 388000 wines and processes over 16000 wines per year. All wine tastings are conducted in private, controlled conditions. The identity of each of the wines is obfuscated. Only the vintage and general type of wine are revealed to the reviewer before tasting [3]. This method of blind tasting removes external factors such as cost or bias that could negatively impact the quality of the reviews. Below is an example of a typical wine review from *Wine Spectator*.

Kosta Browne Pinot Noir Sonoma Coast 2009 95pts

Ripe and deeply flavored, concentrated, and well-structured, this full-bodied red offers a complex mix of black cherry, wild berry and raspberry fruit that's pure and persistent, ending with a pebbly note and firm tannins.

Wine reviews follow a standard 100-point scale:

95-100 *Classic: a great wine*

90-94 *Outstanding: a wine of superior character and style*

85-89 *Very good: a wine with special qualities*

80-84 *Good: a solid, well-made wine*

75-79 *Mediocre: a drinkable wine that may have minor flaws*

50-74 *Not recommended*

The most common components of each review are the sensory analysis of the flavors, acidity, tannins, weight, and finish of the wine. While there are tens of thousands of possible descriptive qualities that could exist within a single review, commonalities emerge when examining reviews side-by-side. The presence/absence of a given quality can be expressed with binary states after processing the input wine review with the Computational Wine Wheel. For the above sample wine, the processed review data will be tabulated as: "berry = 0, raspberry = 1, wild berry = 1, tannins high = 1, tannins low = 0, beauty = 0, …" Note that the Computational Wine Wheel considers the semantic and syntactic analysis of the input text data when making determinations of an input wines qualities. In this example, "wild berry" and "raspberry" flavors are present, but "berry" is not.

The Bordeaux dataset is comprised of 14,349 collections of wine rated between 60 and 100 by wine reviewers. However, 10086 of the dataset observations are wines with score category 90- while 4263 of the dataset observations contain wines with score category 90+. The wines have the likelihood of containing certain any of the 986 wine features derived from the Computational Wine Wheel. A one (1) is used to indicate when any of the wine features is present in any of the collected wines. A zero (0) indicates the feature is not in the wine. So, the input dataset for this project is one that the columns are made up of ones(1s) or zeros(0s) except the Score feature which was grouped into two (90+ or 90) as shown in fig II.I below. However, the dataset is huge and exerts high computational cost. But there are techniques [10, 11] that are used to significantly reduce the computation required at query time, such as indexing training examples, but it is out of our concerns in this paper.

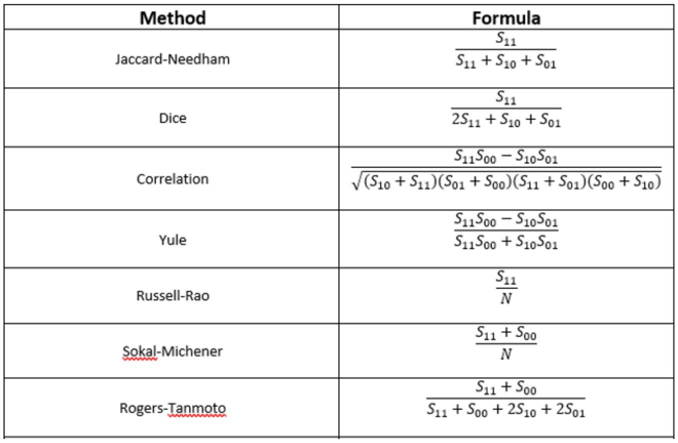
WINE	WET WOOL, WET DOG	ACETIC ACID		ETHANOL	SCORE
WINE A	1	1	…...	0	92
WINE B	0	0	…..	1	87
WINE C	1	1	…..	0	65
WINE D	0	0	…...	0	87

Fig II.I Sample Dataset

3. Methods

3.1 Overview of Distance Metrics

Figure III.1 below defines the distance metrics surveyed. While these formulas have matching dissimilarity indices [5], only the similarity index formulas were evaluated. The term $\mathbf{S_{11}}$ represents intersections between observations in the wine dataset -- instances where wine A and Wine B share an attribute. The terms $\mathbf{S_{01}}$ and $\mathbf{S_{10}}$ represent instances where an attribute is present in one wine and absent in the other. The term $\mathbf{S_{00}}$ represents instances where an attribute is not found in either wine. **N** represents the total number of observations. It is worth noting that there are many similarities between the distance metrics surveyed. The Jaccard-Needham and Dice formulas are very similar and will produce the same accuracy for the wine dataset. Russel-Rao and Sokal-Michener are also very similar, with the latter formula simply considering S_{00} in the numerator.



Method	Formula
Jaccard-Needham	$\frac{S_{11}}{S_{11}+S_{10}+S_{01}}$
Dice	$\frac{S_{11}}{2S_{11}+S_{10}+S_{01}}$
Correlation	$\frac{S_{11}S_{00}-S_{10}S_{01}}{\sqrt{(S_{10}+S_{11})(S_{01}+S_{00})(S_{11}+S_{01})(S_{00}+S_{10})}}$
Yule	$\frac{S_{11}S_{00}-S_{10}S_{01}}{S_{11}S_{00}+S_{10}S_{01}}$
Russell-Rao	$\frac{S_{11}}{N}$
Sokal-Michener	$\frac{S_{11}+S_{00}}{N}$
Rogers-Tanmoto	$\frac{S_{11}+S_{00}}{S_{11}+S_{00}+2S_{10}+2S_{01}}$

Fig III.1: Distance metrics definition.

3.2 K Nearest Neighbor (KNN)

To have a better understanding that goes beyond the idea of determining whether a wine has a score of 90+ or 90-, we found out that it would be best to use the K-NN algorithm because it works efficiently in binary classification. Though KNN has been applied to text categorization since the early days of its research [8]. In pattern recognition, the K-Nearest Neighbor algorithm (KNN) is a method for classifying objects based on the closest training examples in the feature space. KNN is a type of instancebased learning, or lazy learning, where the function is only locally approximated, and all computation is postponed until classification. One of the simplest of all machine learning algorithms is the k-NN algorithm; as the object is categorized by a majority vote of its neighbors and the object is assigned to the most common class among its closest neighbors (k is a positive integer, typically small) [7]. If k is an odd number, it will vote and then simply assigned the object to its closest neighbor's class. That simply justify the reason that even numbers are avoided for the K values, because there are many chances of having ties which will not lead to anything beneficial.

With the use of cross-validation which is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model. In typical cross-validation, the training and validation sets must cross-over in successive rounds such that each data point has a chance of being validated against. One of the basic cross-validation we decided to use is Leave-One-Out cross-validation. Our k-NN algorithm requires that we can calculate the distance between two points using their attributes. The distance formulas are the one that was explained above. The K-Nearest-Neighbours classifier requires storing the entire training set and this is too costly when the training set is large and many researchers have attempted to get rid of the redundancy of the training set to improve this situation [11,12,13,14].

3.3 Evaluation Metrics

As another way to evaluate how effective our classification model is, several standard statistical evaluation metrics are used in this research. We listed all the different parameters that are used and their explanations as well

- True Positive (**TP)**: The real condition is true (1) and predicted as true (1); 90+ wine correctly classified as 90+ wine

- True Negative **(TN**): The real condition is false (-1) and predicted as false (-1); 90- wine correctly classified as 90- wine

- False Positive (**FP**): The real condition is false (-1) but predicted as true (1); 90- wine incorrectly classified as 90+ wine
- False Negative **FN**: The real condition is true (1) but predicted as false (-1); 90+ wine incorrectly classified as 90- wine
- Accuracy: The proportion of wines that has been correctly classified among all wines. Accuracy is a very intuitive metric.

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{1}$$

- Recall: The proportion of 90+ wines was identified correctly. Recall explains the sensitivity of the model to 90+ wine.

$$Recall = \frac{\text{TP}}{\text{TP + FN}} \tag{2}$$

- Precision: The proportion of predicted 90+ wines was correct.

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \tag{3}$$

- F-score: The harmonic mean of recall and precision.

  F-score takes both recall and precision into account, combining them into a single metric.

$$F - score = 2 \times \frac{precision*recall}{(precision+recall)} \tag{4}$$

4. Results

The KNN model was implemented in line with many other statistical evaluation metrics to justify the output of the model. In this research, we calculated the True Positive, False Positive, True Negative, False Negative, Accuracy, Precision, Recall and F-score for the model with unique distance similarity.

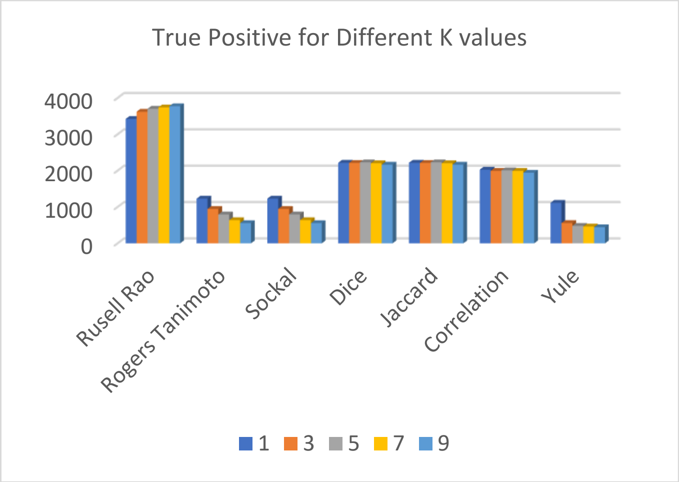


Fig IV.I: True Positive values

The number of nearest neighbors is the list of odd numbers between 1 and 10. Fig IV.I above depicts the True Positive (TP) values. However, Rusell Rao similarity index has the highest True Positive value across the K values. The True Positive value increases as K values increase when KNN is implemented with Rusell Rao similarity index, but reverse is the case when the other distance similarity indices mentioned in this paper are used in the KNN model.

The True Negative values increases across K values and similarity index. Rogers Tanimoto and Sockal similarity indices maintained similar True Negative values across different K values as shown in Fig IV. II.

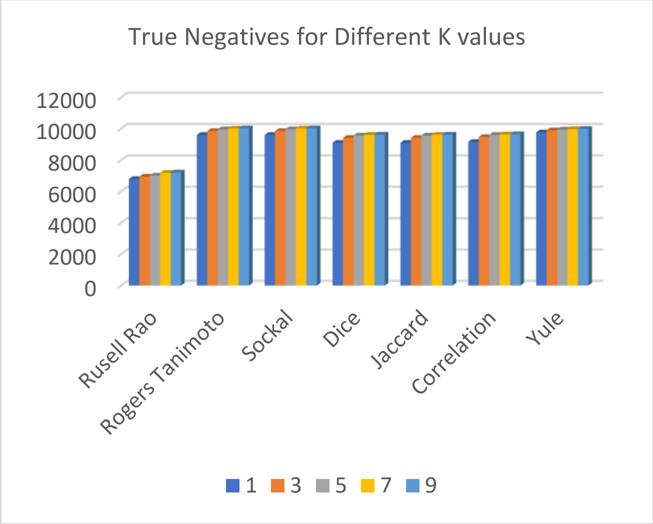


Fig IV. II: True Negative values

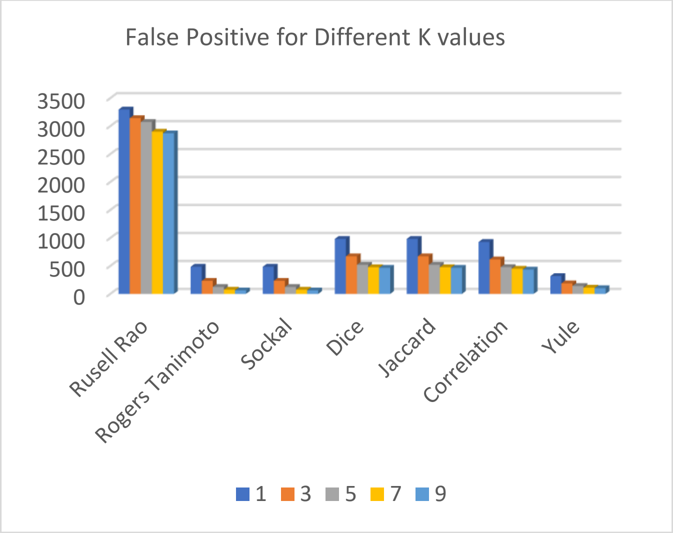


Fig IV. III: False Positive values

The KNN model has highest False Positive values with Rusell Rao similarity index which decreases as the K value increases. The least False Negative values occurred when Rusell Rao similarity index is implemented. Fig IV. IV below shows the False Negative values which increases as K value increases for all similarity indices except for Rusell Rao similarity index where False Negative decreases as K value decreases. The model has high False Negative when implemented with Rogers Tanimoto, Sockal and Yule similarity.

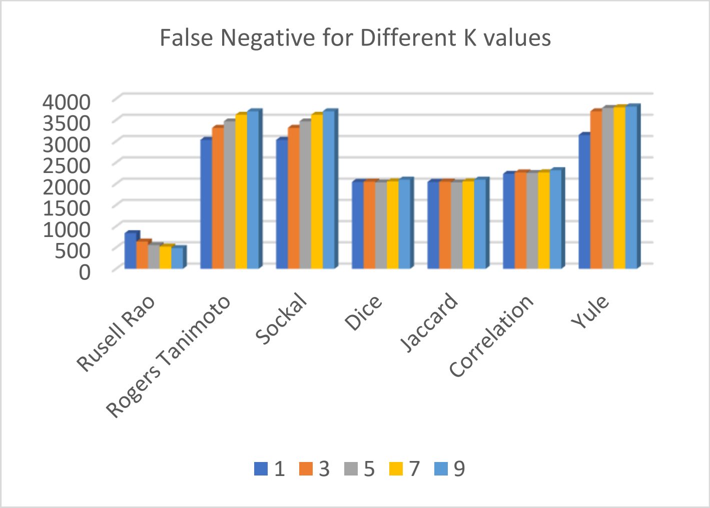


Fig IV. IV: False Negative values

The accuracy of the prediction is showcased in Fig IV.V. Dice and Jaccards similarity indices made predictions with highest accuracy across all values of K. The accuracy of prediction for Yule similarity index decreased as K values increases except when K is 7. However, the accuracy of prediction with Rogers Tanimoto and Sockal similarity indices decreases as K value increases. The accuracy of prediction increases as K value increases for Rusell Rao similarity index. Meanwhile, Dice, Jaccard and Correlation similarity indices increases as K increases from 1 to 7 and decreases when K is 9.

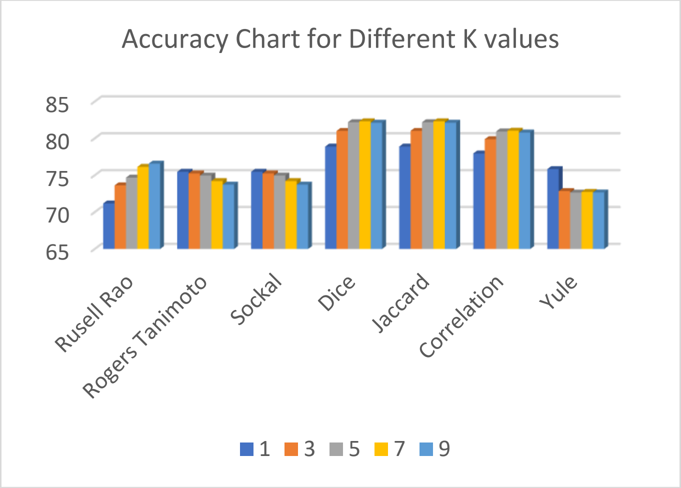


Fig IV.V: Accuracy Values

The Precisions of the model is represented in the table Fig IV.VI. Every similarity index has the least precision when K is 1 except yule similarity index which has the least precision when K is 3 and highest when K is 9. Rusell Rao similarity index has the least precision across all K values and other similarity indices.

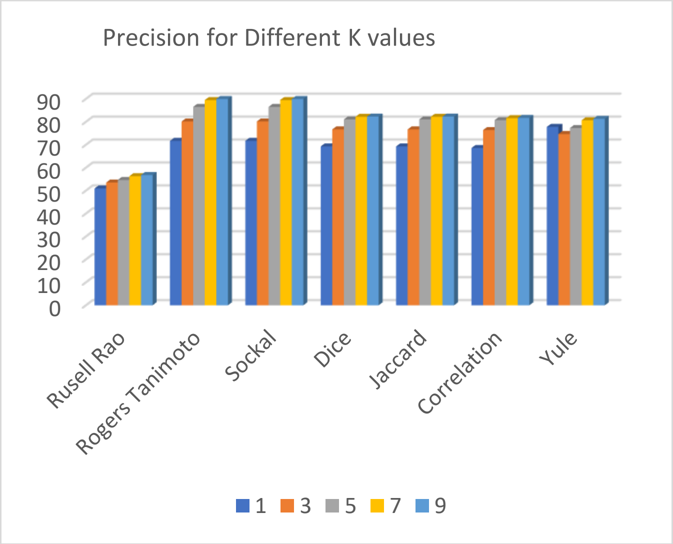


Fig IV.VI: Precision Values

Rusell Rao and Yule similarity indices have the highest and lowest recall values across all values of K respectively as shown in table Fig IV.VII. The recall value for Rusell Rao increases as K value increases. Dice and Jaccards similarity indices maintained the same recall values across K values as Rogers Tanimoto and Sockal similarity indices have the same recall values across K values.

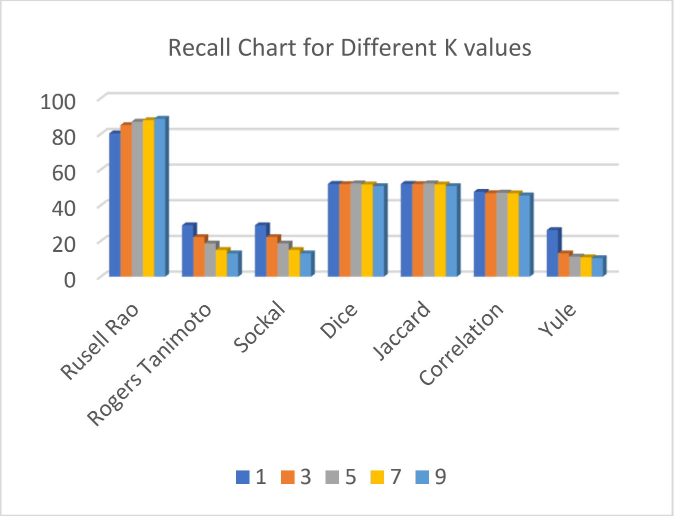


Fig IV.VII: Recall Values

Sockal, Roger Tanimoto similarity and yule similarity index have F-score values that are below 50% while other similarity indices have F-score value above 50%. The rest of the similarity indices seem to be significant for prediction as their F-Score values are above 50%. This is shown in fig IV.VIII below.

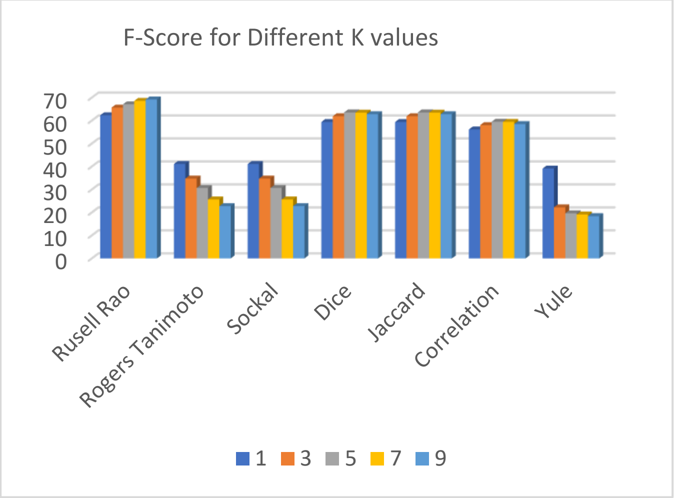


Fig IV.VIII: F-Score Values

5. Conclusion

The main objective of this research is to determine the accuracy of predicting score category of Bordeaux wine reviews. The reviews used in this research were made between year (2010 – 2016) and they were extracted from Wine Spectator. Then, the reviews were transformed from human languages into a more readable computer related form using the Computation Wine Wheel. The dataset used for the research is set of binary values (0s and 1s) as shown in table II.I of wine data section. A value of one (1) indicates that certain specified attribute exists in a certain Bordeaux wine and a value of zero (0) indicates the absence of a wine attribute. A collection of 14,349 wines and 986 different wine attributes define the dataset. Moreover, a list of statistical techniques was implemented to justify the predictive ability of the model. From the results, Jaccard and Dice similarity index have highest accuracy of prediction of 82.326 when K is 7, and their accuracy of prediction increased as K value increased until till K equal to 9. Correlation similarity index exhibited the same trend as Dice and Jaccard's similarity. Sockal and Roger Tanimoto similarity index have highest precision values for all K values. In other words, the model predicts True Positives more accurately when the Rogers Tanimoto and Sockal similarity indices are implemented. The high level of False Positive produced by the model when Russell Rao similarity index was implemented may be attributed to the imbalance nature of the dataset. However, the other similarity indices have a fair share of False Positive. Russell Rao similarity index has the highest recall values across the values of K. Sockal, Roger Tanimoto similarity and yule similarity index have F-score values that are below 50% while other similarity indices have F-score value above 50%.

6. Reference

[1] Chen, Dr. Bernard, Rhodes, C., Crawford, A., Hambuchen, L. Wine Informatics: Applying Data Mining on Wine Sensory, Accepted by 2014 Workshop on Domain Driven Data Mining (DDDM 2014).

[2] Wine Spectator's 100-Point Scale | Wine Spectator, winespectator.com, 2019.Available: https://www.winespectator.com/articles/scoring-scale

[3] B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen, "Wineinformatics: Applying Data Mining on Wine Sensory Reviews Processed by the Computational Wine Wheel," *2014 IEEE International Conference on Data Mining Workshop*, Shenzhen, 2014, pp. 142-149

[4] Wine Spectator Home |Wine Spectator. Available at: https://www.winespectator.com/

[7] Wang, H.: Nearest Neighbours without k: A Classification Formalism based on Probability, technical report, Faculty of Informatics, University of Ulster, N. Ireland, UK (2002)

[8] Sebastiani, F.: Machine Learning in Automated Text Categorization. ACM Computing Surveys 34(1), 1–47 (2002)

[9] Wilson, D.R., Martinez, T.R.: Reduction Techniques for Exemplar-Based Learning Algorithms. Machine learning 38(3), 257–286 (2000)

[10] Mitchell, T.: Machine Learning. MITPress/McGraw-Hill (1997)

[11] Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, UK

[11] Hart, P.: The Condensed Nearest Neighbor Rule. IEEE Transactions on Information Theory 14, 515–516 (1968)

[12] Gates, G.: The Reduced Nearest Neighbor Rule. IEEE Transactions on Information Theory 18, 431–433 (1972)

[13] Alpaydin, E.: Voting Over Multiple Condensed Nearest Neighbors. Artificial Intelligence Review 11, 115–132 (1997); Kluwer Academic Publishers (1997)

[14] Kubat, M., Jr., M.: Voting Nearest-Neighbor Subclassifiers. In: Proceedings of the 17th International Conference on Machine Learning, ICML 2000, pp. 503–510, Stanford, CA, June 29-July 2 (2000)