Statistical Significance and Utility of Data-Driven Functional Dependencies of Wine Quality Data of Numerical Attributes

HYONTAI SUG
Department of Computer Engineering,
Dongseo University,
47 Jurye-ro, Sasang-gu, Busan, 47011,
REPUBLIC OF KOREA

Abstract: - There has been a lot of research work to find out functional dependencies algorithmically from databases. But, when the databases consist of numerical attributes, some of the found functional dependencies might not be real functional dependencies, because numerical attributes can have a variety of values. On the other hand, regression analysis is an analysis method in which a model of the observed continuous or numerical variables is obtained and the degree of fit is measured. In this paper, we show how we can determine whether the found functional dependencies of numerical attributes have explanatory power by doing multivariate linear regression tests. We can check their explanatory power by way of adjusted R-squared, as well as other statistics like multicollinearity, the Durbin-Watson test for independence, and the F value for suitability of the regression models. For the experiment, we used the wine quality data set of Vinho Verde in the UCI machine learning library, and we found out that only 48.7% and 30.7% of functional dependencies found by the algorithm called FDtool have explanatory power for the red wine and white wine data set respectively. So, we can conclude that we should be careful when we want to apply the functional dependencies found by the algorithm. In addition, as a possible application of the found functional dependencies in the conditional attributes of the data sets, we have generated a series of random forests by dropping redundant attributes that appear on the right-hand side of the explanatory functional dependencies and acquired good results. So, we can also conclude that we may reduce our efforts by not collecting the data of the redundant attribute to check the wine quality because we can use samples with as few attribute values as possible in mass-produced wines like Vinho Verde.

Key-Words: - Databases, machine learning, classification, regression, numerical attributes, functional dependency, wine quality, preprocessing, relations

Received: July 17, 2022. Revised: August 17, 2023. Accepted: September 21, 2023. Published: October 10, 2023.

1 Introduction

Vinho Verde, literally ‘green wine’, is a Portuguese wine produced mainly in the region around the Minho River in northwestern Portugal. Vinho Verde has nothing to do with the grape variety, and it means young wine, which is the opposite of aged wine and is consumed within one year of being bottled. Wine certification and quality assessment are key elements in the Portuguese wine industry, [1]. Certification prevents the illegal adulteration of wines and assures quality for the wine market. Quality assessment is often part of the certification process and can be used to improve winemaking, for example, by identifying the most important factors. Wine quality assessment is generally assessed by physicochemical and sensory tests. Physicochemical tests are used to characterize wine to several physicochemical factors, while sensory tests rely on human experts, and determining wine quality based on data has attracted many research interests, [2], [3], since the wine quality data set was open to the public without relevant attribute selection, [4].

On the other hand, assessing functional dependency is an important step for the normalization of relational databases. When we check functional dependencies for a relation, we check whether each set of attributes has a many-to-one relationship with all the values that may appear in the attributes. And such a decision is made by the database designers, [5]. As a way of assisting database designers, there is a series of research work to find out such functional dependencies algorithmically from data, so several efficient algorithms have been suggested, [6]. But, some functional dependencies found by data may not be real functional dependencies especially when we have numerical or continuous attributes. Note that the two words, numerical and continuous, are used interchangeably. For example, suppose we have a wine relation like Table 1. There are three attributes in the table; alcohol, chlorides, and density. Alcohol sweetens the wine and affects its body feel, chlorides dictate the saltiness and acidity of the wine, and density determines the body feel of the wine.

Table 1. A wine relation

Alcohol	Chlorides	Density
8.8	0.045	1.001
9.5	0.049	0.994
10.1	0.05	0.9951
9.9	0.058	0.9956

Then, we have functional dependencies of one-to-one based on Table 1 as follows;

{alcohol} -> {chlorides}, {chlorides}->{alcohol}.
{alcohol} -> {density}, {density}->{alcohol},
{chlorides} -> {density}, {density}->{chlorides}
{alcohol, chlorides} -> {density}, and so on.

Even though the above examples are some extremes, we can see the fact that we may have many functional dependencies that may not be real functional dependencies if we decide on functional dependencies by stored data only.

On the other hand, a statistical method to check dependency between numerical attributes is linear regression or multiple regression, depending on whether the left and right-hand side of the functional dependency consists of one or several attributes or variables, [7]. We want to check dependency between attributes in the functional dependencies by regression test, and this approach is very novel in the research area of functional dependency by data.

Therefore, we want to select functional dependencies of the wine quality data set by doing the regression test for the same attribute sets which are found to have functional dependencies by data algorithmically, and also select attributes on the right-hand side in the functional dependencies as redundant attributes. As a possible application, we may use the found fact to generate machine learning models without using the redundant attribute to check wine quality, because acquiring samples with as few attribute values as possible is important to reduce our efforts in mass-produced wines like Vinho Verde.

2 Related Work

Functional dependencies are used to determine the many-to-one relationship of values between the attributes of relations or relation variables, [8], and serve as a criterion for judgment for normalization. Several algorithms have been proposed to discover functional dependencies based on data stored in relations. The algorithms try to find functional dependencies as efficiently as possible, so several efficient algorithms have been suggested, [9], [10], [11].

Because functional dependencies represent many-to-one correspondences of attribute values including one-to-one relationships of values that appear in a relation, if there is a functional dependency, there may be some statistical dependency between the two sets of attributes that appear in the left-hand side (LHS) and the right-hand side (RHS) of the functional dependency. When we have continuous variables or attributes consisting of each RHS and LHS of length one, linear regression can be used to determine statistical significance. When we have continuous variables or attributes consisting of LHS of length more than one and RHS of length one, multivariate simple linear regression can be used to determine statistical significance, [12].

Random forest generates many decision trees based on random sampling with replacement and uses the many decision trees to classify, [13]. When each decision tree is generated, a random selection of root attributes for each subtree and no pruning is performed, and classes are determined by majority vote by decision trees in the forest so that overfitting can be avoided. Random forest is known as one of the most reasonable machine-learning algorithms across a wide range of data, [14].

There are some public wine data sets available and a lot of related research work done. For example, in 1991 a data set called ‘wine’ was loaded into the UCI machine learning repository, [15]. The data set contains 178 instances consisting of 13 chemical constituents to classify three wine cultivars from Italy. Statistical inference was considered for variables of interest when auxiliary variables are observed along with the variables of interest that do not have enough data, [16]. More recently, neural network and support vector machine classifiers were made for small syntactic wine data having 173 instances with 13 different physiochemical characters to generate accurate classifiers, [17]. Another relatively large data set called ‘wine quality’, which is also our target data set, was loaded in 2009 in the UCI machine learning repository, [4]. Using the data set, a decision tree was generated to assess the wine quality and achieved an accuracy of 60%, [18]. An over-sampling technique was used to surmount the data imbalance problem of the data set, [19]. While the original paper using the data set used SVM and achieved an accuracy of 62~ 64%, [4], regression analysis was performed to classify the red wine data set, and the area under the ROC curve of the best-fit model was obtained as 0.73, [20]. Because the wine data set has some redundant attributes, a Pearson correlation coefficient-based table, called a heat map, was used to remove some redundant attributes and shows good accuracies, especially in the random forest among four different classifiers, [21]. By permuting attributes of the white wine data set and other three different data sets, the mean decrease in accuracy or how the prediction gets worse is calculated, and some reduced attribute sets showed better results for SVM and random forest for root mean squared error and Pearson correlation coefficient, [22].

3 Problem Formulation

The definition of functional dependency based on data can be defined as follows, [8].

Definition 1. Let **r** be a relation over the set of attributes U and A, B be any subset of U. Then B is functionally dependent on A, A → B, if and only if each A value in **r** is associated with precisely one B value. □

Note that in statistics the term 'variable' is used instead of the term 'attribute' in the database.

The wine data set has eleven continuous or numerical conditional attributes and one decisional attribute that classifies wine quality into eleven classes. We want to find statistically significant functional dependencies between conditional attributes and find redundancy in the conditional attributes.

When we try to find functional dependencies based on data, relatively much more functional dependencies could be found if the size of the data is not large enough and attributes or variables are continuous. Moreover, some of the found functional dependencies may not be statistically significant functional dependencies because of the property of functional dependency of many-to-one relationships. And, because continuous variables can have much more variety of values, there can be a high possibility of a many-to-one or one-to-one relationship between attribute values. Therefore, we want to do statistical tests for the found functional dependencies based on data, especially for the data set consisting of continuous attributes like the wine data set.

The functional dependencies we want to analyze are that both the independent variable and the dependent variable are continuous, and the independent variables are multiple, so multivariate linear regression analysis can be performed. To do multivariate linear regression analysis, there are several statistics we should consider: the coefficient of determination (adjusted R-squared), and multicollinearity, [12], [23].

The coefficient of determination or R-squared is an indicator of how well the independent variable explains the dependent variable in the regression model. It is also called explanatory power, because the higher the coefficient of determination, the more the independent variable explains the dependent variable. In addition, as the number of independent variables increases, the R-squared also increases, but it tends to increase even if variables that do not explain the dependent variable well are added, so the adjusted R-squared is used. In general, it is believed that the coefficient of determination should exceed 20%.

Multicollinearity refers to the inclusion of variables that are too similar among the independent variables. In other words, multicollinearity shows a strong correlation between independent variables in regression analysis. Usually, if the variance inflation factor (VIF) is 10 or more, there is a strong correlation between the independent variables. A simple method to solve the multicollinearity problem is removing one or several of the highly correlated independent variables, [24].

The other statistics are the Durbin-Watson test for independence test and the F value in the analysis of variance for suitability of the regression model. The independence test of the residuals in regression analysis is performed by the Durbin-Watson test that is denoted by d. The d value has a value between 0 and 4, and the closer the d value is to 2, the less auto-correlated it is. Auto-correlation is more likely to occur in time-series data. In the analysis of variance, if the significance probability for the F test is less than significance level 0.05, we can consider the regression model suitable.

A recent study implemented an automated program called FDtool that finds functional dependencies in the datasets of tabular form, [25]. FDTool is a Python-based open-source program to mine functional dependencies and candidate keys algorithmically. The number of attributes of the input data is limited to 26. For the statistical test, we will use a well-known tool, IBM SPSS, [26], for our experiment. To generate random forest we will use an open-source tool called Weka, [27].

3.1 Experimental Procedure

We want to check whether the wine dataset consisting of conditional and decisional attributes for data mining has statistically significant functional dependencies within the conditional attributes. To find functional dependencies, we use

FDtool, then, if some functional dependencies are found, we try to do statistical tests. After the statistical tests, we select functional dependencies having explanatory power.

Since the attribute values of the right-hand side are determined by the attribute values of the left-hand side because of the many-to-one property of functional dependency, we can consider the attributes on the right-hand side may be redundant.

As a possible application of the found statistical facts of the functional dependencies of explanatory power, we try to generate several random forests by dropping the redundant attributes.

4 Experimentation

There are two wine data sets available in the UCI machine learning repository, [28]. The 'Wine' data set has 178 instances, while the 'Wine quality' data set has up to 4,898 instances. We chose the 'wine quality' data set, [4], for our experiments, because the data set is well-known and contains a large amount of data records with numerical attributes that are a very good fit for our experiment. The data set contains two wine quality data sets, related to red and white vinho wine samples from the north of Portugal. The goal is to model wine quality based on physicochemical tests. The datasets consist of eleven conditional attributes and one decisional attribute and no dependency tests between the conditional attributes were conducted in the original data set.

4.1 Red Wine Dataset

The data set has 1,599 records and has 11 conditional attributes and one decisional attribute, named 'quality'. The 11 conditional attributes consist of numerical attributes as in Table 2.

Table 2. The Property of attributes of the red wine data set

Attribute	Value Range	Distinct values	Mean	Standard Dev.
Fixed acidity	4.6 ~ 15.9	96	8.32	1.741
Volatile acidity	0.12 ~ 1.58	143	0.528	0.179
Citric acid	0 ~ 1	80	0.271	0.195
Residual sugar	0.9 ~ 15.5	91	2.539	1.41
Chlorides	0.012 ~ 0.611	153	0.087	0.047
Free sulfur dioxide	1 ~ 72	60	15.875	10.46
Total sulfur dioxide	6 ~ 289	144	46.468	32.895
Density	0.99 ~ 1.004	436	0.997	0.002
pH	2.74 ~ 4.01	89	3.311	0.154
Sulphate	0.33 ~ 2	96	0.658	0.17
Alcohol	8.4 ~ 14.9	65	10.423	1.066
Quality	11 nominal values (0 ~ 10)			

Fixed acidity represents the acidity of the wine, volatile acidity is associated with the aroma of the wine, and citric acid plays a role in keeping it fresh, and is associated with acidification. The residual sugar enhances the sweetness, and chloride dictates the saltiness and acidity of the wine. Free sulfur dioxide, total sulfur dioxide, and sulfate are sulfur compounds that kill certain bacteria and yeasts, increasing the preservation of wine. Density determines the body feel of the wine, and pH indicates the degree of acidity. Alcohol sweetens the wine and affects its body feel. A quality score by wine experts was assigned which graded each wine sample with a score ranging from 0 to 10.

4.1.1 Checking Functional Dependencies for the Red Wine Data Set

FDtool was used to find functional dependencies (FDs) in the eleven conditional attributes. A total of 618 functional dependencies were found. The found FDs have four or five attributes on the left-hand side (LHS) of the FDs and one attribute on the right-hand side (RHS) of the FDs. The following are some examples:

{density, sulphates, total_sulfur_dioxide} -> {free_sulfur_dioxide}
{alcohol, chlorides, sulphates, fixed_acidity} -> {density}

. . . . . .
. . . . . .
{alcohol, free_sulfur_dioxide, residual_sugar, total_sulfur_dioxide, citric_acid} -> { pH}
{alcohol, free_sulfur_dioxide, residual_sugar, total_sulfur_dioxide, citric_acid} -> {chorides}

Table 3 shows the total number of functional dependencies for each attribute on the right-hand side of the found functional dependencies based on data.

Table 3. The number of functional dependencies found for each attribute in the right-hand side of the red wine data set

RHS attribute of functional dependencies	**The total number of functional dependencies found**
Density	108
Citric_acid	86
pH	77
Free_sulfur_dioxide	38
Volatile_acidity	74
Chloride	91
Sulphate	61
Residual_sugar	83
TOTAL	618

Because FDtool finds simple many-to-one relationships in attribute values, the found functional dependencies may not be statistically significant. Therefore, multivariate linear regression analysis was performed for each found functional dependency. The R-squared value can be an indication of how much the independent variable explains the dependent variable in the regression model. In general, it should exceed 20%, and the higher this number, the higher the explanatory power. However, this value increases even with the addition of a dependent variable that does not explain the dependent variable well, so an adjusted R-squared value is used. 301 functional dependencies have adjusted R-squared values exceeding 20%. So, only 48.7% of found functional dependencies based on data have explanatory power. All the d values of the 301 FDs have a value between 1.272 and 1.797, so we can consider they are close to the d value of 2, which means they are less auto-correlated. In the analysis of variance, all the significance probabilities for the F test are 0.0, and since they are less than the significance level of 0.05, we can consider all the regression models of the 301 FDs are suitable. Table 4 shows the summary.

Table 4. The distribution of the number of FDs having adjusted R-squared values of 20% or more for the red wine data set

RHS attribute of FDs	**The number of FDs having adjusted R-squared value of 20% ~ 50%**	**The number of FDs having adjusted R-squared value of 50% ~ 82%**
Density	47	39
Citric_acid	52	26
pH	31	23
Free_sulfur_dioxide	23	0
Volatile_acidity	23	0
Chloride	19	1
Sulphate	13	0
Residual_sugar	4	0
TOTAL	212	89

The following Table 5 (Appendix) shows the combined result of Table 3 and Table 4.

Multicollinearity was also considered for FDs having adjusted R-squared value of 20% or more. Multicollinearity occurs when highly correlated independent variables are included in the model at the same time. If the variance inflation factor (VIF) is more than or equal to 10, highly correlated independent variables exist. All the values of VIF are in the range of 1.001 ~5.040. So, all the FDs having adjusted R-squared value of 20% or more are good for multicollinearity.

4.1.2 Generating Random Forest

As a possible application of found explanatory functional dependencies, several random forests are generated with 1000 random trees in the forest with 10-fold cross-validation. The accuracy of the random forest with the original red wine data set is 70.5441% which is better than 62.4% of the SVM used in the original paper in 5-fold cross-validation, [4]. The accuracy of the random forest with the original red wine data set is 70.3565% in 5-fold cross-validation. Because the attributes in the RHS of found FDs have a high possibility of redundancy, we tried to generate random forests by dropping such attributes. By dropping each attribute in the RHS of FDs one by one based on the descending order of the number of FDs having adjusted R-squared values of 20% or more of the red wine dataset, the random forest algorithm was generated. Table 6 shows the accuracies of the random forests. All experiments are performed in 10-fold cross-validation.

Table 6. Random forests of the red wine data set by dropping some part of attributes

Dropped attributes	**Accuracy**
Nothing (original data)	70.5441%
Density	70.4816%
Density, citric_acid	70.8568%
Density, citric_acid, pH	70.9819%
Density, citric_acid, pH, free_sulfur_dioxide	**71.2320%**
Density, citric_acid, pH, free_sulfur_dioxide, volatile_acidity	70.4816%
Density, citric_acid, pH, free_sulfur_dioxide, volatile_acidity, chloride	68.7930%
Density, citric_acid, pH, free_sulfur_dioxide, volatile_acidity, chloride, sulphate	66.9794%
Density, citric_acid, pH, free_sulfur_dioxide, volatile_acidity, chloride, sulphate, residual_sugar	65.0407%

Note that dropping attributes like density, citric_acid, pH, and free_sulfur_dioxide achieves better accuracy compared to the original data set in which no attributes are dropped. So we may not collect these values anymore, as a result, saving our efforts and resources for us to collect these values.

4.2 White wine dataset

The data set has 4,898 records and has 11 conditional attributes and one decisional attribute, named ‘quality’. The 11 conditional attributes consist of numerical attributes as in Table 7 (Appendix).

4.2.1 Checking functional dependencies for the white wine dataset

FDtool was used to find functional dependencies (FDs) in the eleven conditional attributes. Because the size of the white wine data set is more than three times the size of the red wine data set, we have more diversity in attribute values, so only a total of 390 functional dependencies were found. Remember that we have 618 FDs for the red wine data set. The found FDs have four to six attributes on the left-hand side (LHS) of the FDs and one attribute on the right-hand side (RHS) of the FDs. The following are some examples:

{alcohol, chlorides, density, free_sulfur_dioxide} -> {sulphates}

{chlorides, density, fixed_acidity, residual_sugar} -> {sulphates}

. . . . . .

. . . . . .

{alcohol, chlorides, fixed_acidity, free_sulfur_dioxide, residual_sugar} -> {sulphates}

{alcohol, density, sulphates, fixed_acidity, volatile_acidity} -> { citric_acid}

. . . . . .

. . . . . .

{alcohol, density, fixed_acidity, free_sulfur_dioxide, volatile_acidity, pH} -> {chlorides}

{alcohol, density, fixed_acidity, volatile_acidity, pH, citric_acid} -> {sulphates}

Table 8 shows the number of functional dependencies for each attribute on the right-hand side of the found functional dependencies based on data.

Table 8. The number of functional dependencies found for each attribute on the right-hand side of the white wine data set

RHS attribute of functional dependencies	**The total number of functional dependencies found**
Density	63
Citric_acid	34
pH	64
Free_sulfur_dioxide	55
Volatile_acidity	35
Chloride	47
Sulphate	92
TOTAL	390

Because the wine data set consists of continuous conditional attributes, we have more diversity in the values of attributes, and the white wine data set has about three times more data than the red wine data set so we have a smaller number of functional dependencies as in the above Table 8. One difference with the red wine data set is that no functional dependencies are found for the attribute ‘residual_sugar’ when it is RHS of functional dependency in the white wine data set, while 83 functional dependencies are found in the red wine data set. Because the found functional dependencies may not be statistically significant, multivariate linear regression analysis was performed for each found functional dependency like the case of the red wine data set. 122 functional dependencies have adjusted R-squared values exceeding 20%. So, only 30.7% of found functional dependencies have explanatory power. All the d values of the 122 FDs have a value between 1.364 and 1.749, so we can consider they are closer to the d value of 2, which means they are less auto-correlated. In the analysis of variance, all the significance probabilities for the F test are 0.0, and since they are less than the significance level of 0.05, we can consider all the regression models of the 122 FDs are suitable.

Table 9. The distribution of the number of FDs having adjusted R-squared values of 20% or more for the white wine data set

RHS attribute of FDs	**The number of FDs having adjusted R-squared value of 20% ~ 50%**	**The number of FDs having adjusted R-squared value of 50% ~ 82%**
Density	12	49
Citric_acid	0	0
pH	19	1
Free_sulfur_dioxide	41	0
Volatile_acidity	0	0
Chloride	0	0
Sulphate	0	0
TOTAL	72	50

The following Table 10 (Appendix) shows the combined result of Table 8 and Table 9.

Note that we have no explanatory FDs for RHS attributes, citric_acid, volatile_acidity, chlorides, and sulphate for the white wine data set. Multicollinearity was also considered for FDs having adjusted R-squared value of 20% or more. All the VIF are in the range of 1.004 ~4.552 except one FD, {alcohol, density, sulphate, fixed_acidity, residual_sugar, volatile_acidity} -> {pH} with adjusted R-squared value of 52.6%, d is 1.587, and VIF for each attribute in the LHS is as the following Table 11.

Table 11. The VIF (variance inflation factor) for each attribute in the LHS of an FD, {alcohol, density, sulphate, fixed_acidity, residual_sugar, volatile_acidity} -> {pH}

Attribute	**VIF**
Alcohol	5.109
Density	**16.004**
Sulphate	1.121
Fixed_acidity	1.399
Residual_sugar	**7.137**
Volatile_acidity	1.03

So, all the FDs having adjusted R-squared value of 20% or more except the one above are good with multicollinearity.

4.2.2 Generating Random Forest

Several random forests are generated with 1000 random trees in the forest with 10-fold cross-validation. The accuracy of the random forest for the original white wine data set is 70.4777% which is slightly better than 64.6% of the SVM used in the original paper in 5-fold cross-validation, [4]. The accuracy of the random forest for the white wine data set is 69.1% in 5-fold cross-validation. Because the attributes in the RHS of found FDs have a high possibility of redundancy, we tried to generate random forests by dropping such attributes. By dropping each attribute in the RHS of FDs one by one based on the descending order of the number of FDs having adjusted R-squared values of 20% or more of the white wine dataset, the random forest algorithm was applied. Table 12 shows the accuracies of the random forests. All experiments are performed in 10-fold cross-validation.

Table 12. Random forests by dropping some attributes

Dropped attributes	**Accuracy**
Nothing (original data)	70.4777%
Density	70.4369%
Density, pH	69.9265%
Density, pH, free_sulfur_dioxide	69.7019%

Note that dropping attributes like density achieves similar accuracy to the original data set in which no attributes are dropped. Additionally, because density and residual_sugar have relatively high VIF as in Table 11, we tried to drop the two attributes, density and residual_sugar, to solve the multicollinearity problem. A random forest with an accuracy of **70.7023%** was generated, and this accuracy is slightly better than the accuracy of 70.4777% from the original data. So, we may not need to collect data for attributes like density and residual_sugar for the white wine data set, thus we could save our efforts in data collection.

5 Conclusion

A lot of research work to find out functional dependencies algorithmically from data has been published, and several efficient algorithms have been suggested. But, when we have continuous or numerical attributes, it is highly possible that some parts of functional dependencies found by the algorithms may not be real functional dependencies, because continuous attributes can have a variety of values, and because a functional dependency has to satisfy many-to-one relationship between attributes values only, and the algorithms do not check all the possible attribute values that are not present in the data. In this paper, we show how we can determine whether the found functional dependencies have the statistical significance of explanatory power by doing a multivariate linear regression test for each algorithmically found functional dependency to compensate for the weakness of the algorithmic methods. We can check their explanatory power by calculating adjusted R-squared, and we also considered other statistics like multicollinearity, the Durbin-Watson test for independence, and the F value for suitability of the regression model. For our experiment, we used the wine quality data set of Vinho Verde in the UCI machine learning library, and we found that only 48.7% of functional dependencies found by the algorithm called FDtool have explanatory power for the red wine data set, while only 30.7% of functional dependencies have explanatory power for the white wine data set. From these findings, we can conclude that we should be careful when we want to take advantage of the functional dependencies found by the algorithm because the algorithm finds too much functional dependency on numerical attributes.

In addition, as a possible application of found explanatory functional dependencies in the conditional attributes, we have generated random forests by dropping redundant attributes that appear at the right-hand side of the explanatory functional dependencies and acquired good results. So, we can also conclude that we can reduce our efforts by not collecting redundant attribute values to check wine quality because we can use samples of as few attribute values as possible in mass-produced wines like Vinho Verde. Further developments for more practical applications can be the discretization of numerical attributes to reduce the number of found functional dependencies by the algorithm, and the task can be challenging because there are many discretization methods available, and selecting the best one is difficult.

Acknowledgment:
This work was supported by Dongseo University, "Dongseo Frontier Project" Research Fund of 2022.

References:

[1] S.E. Ebeler, Linking Flavor Chemistry to Sensory Analysis of Wine. In: Teranishi, R., Wick, E.L., Hornstein, I. (eds) *Flavor Chemistry*. Springer, Boston, MA., 1999. https://doi.org/10.1007/978-1-4615-4693-1_35.

[2] C.E. Butzke, S.E. Ebeler, Survey of analytical method and winery laboratory proficiency, *American Journal of Enology and Viticulture*, Vol.50, pp.461-465, DOI: 10.5344/ajev.1999.50.4.461.

[3] K.R. Dahal, J.N. Dahal, H. Banjade, S. Gaire, Prediction of wine quality using machine learning algorithms, *Open Journal of Statistics*, Vol.11, No.2, 2021, pp.278-289, DOI: 10.4236/ojs.2021.112015.

[4] P. Cortez, A. Cerderia, F. Almeida, T. Matos, J. Reis, Modeling wine preferences by data mining from physicochemical properties, *Decision Support Systems*, Vol. 47, Issue 4, 2009, pp.547-553.

[5] C.J. Date. *Database Design and Relational Theory: Normal Forms and All That Jazz*, 2nd ed., Apress, 2019.

[6] N. Asghar, A. Ghenai, *Automatic Discovery of Functional Dependencies and Conditional Functional Dependencies: A Comparative Study*, University of Waterloo, April 2015.

[7] T.Z. Keith, *Multiple Regression and Beyond: An Introduction to Multiple Regression and Structural Equation Modeling*, 3rd ed., Routledge, 2019.

[8] C. J. Date, *An Introduction to Database Systems*, 8th ed., Pearson, 2003.

[9] N. Asghar, A. Ghenai, *Automatic Discovery of Functional Dependencies and Conditional Functional Dependencies: A Comparative Study*, University of Waterloo, April 2015

[10] L. Caruccio, S. Cirillo, V. Deufemia, and G. Polese, Incremental Discovery of Functional Dependencies with a Bit-vector Algorithm, *Proceedings of the 27th Italian Symposium on Advanced Database Systems*, 2019, pp.146-157.

[11] J. Liu, J. Li, C. Liu, and Y. Chen, Discover dependencies from data – a review, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 24, No. 2, 2012, pp.251-264.

[12] D.C. Montgomery, E.A. Peck, G.G. Vining, *Introduction to Linear Regression Analysis*, 5th ed., Willey, 2012.

[13] L. Breiman, Random Forests, *Machine Learning*, Vol.45, No.1, pp.5-32, 2001.

[14] A. Lulli, L. Oneto, D. Anguita, Mining Big Data with Random Forests, *Cognitive Computation*, Vol.11, pp.294-316, 2019.

[15] S. Aeberhard, M. Forina, Wine, *UCI Machine Learning Repository*, 1991, DOI: https://doi.org/10.24432/C5PC7J.

[16] S. Imori, H. Shimodaira, An Information Criterion for Auxiliary Variable Selection in Incomplete Data Analysis, *Entropy*, 2019, DOI: 10.3390/e21030281.

[17] D.K. Jana, P. Bhunia, S.D. Adhikary, A. Mishra, Analyzing of salient features and classification of wine type based on quality through various neural network and support vector machine classifiers, *Results in Control and Optimization*, Vol.11, 2023, DOI: https://doi.org/10.1016/j.rico.2023.100219.

[18] S. Lee, J. Park, K. Kang, Assessing wine quality using a decision tree, *2015 IEEE International Symposium on Systems Engineering*, 2015, DOI: 10.1109/SysEng.2015.7302752.

[19] G. Hu, T. Xi, F. Mohammed, H. Miao, Classification of wine quality with imbalanced data, *2016 IEEE International Conference on Industrial Technology*, 2016, DOI: 10.1109/ICIT.2016.7475021.

[20] A. Rajini, V.S.H. Peyyeti, A.S. Goteti, Selection of significant features and prediction of red wine quality using logistic regression, *AIP Conference Proceedings 2707*, 040015, 2023, DOI: https://doi.org/10.1063/5.0146762.

[21] P. Dhaliwal, S. Sharma, L. Chauhan, Detailed study of wine dataset and its optimization, *International Journal of Intelligent Systems and Applications*, 2022, 5, pp.35-46, DOI: 10.5815/ijisa.2022.05.24.

[22] C. Dewi, R. Chen, Random forest and support vector machine on features selection for regression analysis, *International Journal of Innovative Computing, Information and Control*, Vol.15, No 6, 2019, pp.2027-2037.

[23] Tutoraspire.com. *How to test for multicollinearity in SPSS*, https://www.statisticalpoint.com/multicollinearity-spss [Accessed on 27/09/2023]

[24] C.F. Dormann, J. Elith, S. Bacher, C. Buchmann, G. Carl, G. Carré, J.R.G. MarquéZ, B. Gruber, B. Lafourcade, P.J. Leitão, T. Münkemüller, C. McClean, P.E. Osborne, B. Reineking, B. Schröder, A.K. Skidmore, D. Zurell, S. Lautenbach, Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, *Ecography*, Vol.36, Issue 1,2013, pp.27-46.

[25] M. Buranosky, E. Stellnberger, E. Pfaff, D. Diaz-Sanchez, C. Ward-Caviness, *FDTool: a Python application to mine for functional dependencies and candidate keys in tabular form* [version 2, peer review: 2 approved], F1000Research 2019, 7:1667, https://doi.org/10.12688/f1000research.16483.2.

[26] A. Field, *Discovering Statistics Using IBM SPSS Statistics: North American Edition*, 5th ed., SAGE Publications Ltd., 2017.

[27] E. Frank, M.A. Hall, I.H. Witten, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Fourth Edition, 2016.

[28] M. Lelly, R. Longjohn, K. Nottingham, The UCI Machine Learning Repository, https://arcchive.ics.uci.edu [Accessed on 27/09/2023]

APPENDIX

Table 5. The number of functional dependencies for each attribute on the right-hand side of the red wine data set

RHS attribute of FDs	**The number of FDs having adjusted R-squared value of less than 20%**	**The number of FDs having adjusted R-squared value of 20% ~ 50%**	**The number of FDs having adjusted R-squared value of 50% ~ 82%**	**Total & explanatory FD ratio**
Density	22	47	39	108 (79.6%)
Citric_acid	8	52	26	86 (90.7%)
pH	23	31	23	77 (70.1%)
Free_sulfur_dioxide	15	23	0	38 (60.5%)
Volatile_acidity	51	23	0	74 (31.1%)
Chloride	71	19	1	91 (22%)
Sulphate	48	13	0	61 (21.3%)
Residual_sugar	79	4	0	83 (4.8%)
TOTAL	317	212	89	618 (48.7%)

Table 7. The properties of the white wine data set

Attribute	**Value Range**	**Distinct values**	**Mean**	**Standard Dev.**
Fixed acidity	3.8 ~ 14.2	68	6.855	0.844
Volatile acidity	0.08 ~ 1.1	125	0.278	0.101
Citric acid	0 ~ 1.66	87	0.334	0.121
Residual sugar	0.6 ~ 65.8	310	6.391	5.072
Chlorides	0.09 ~ 0.346	160	0.046	0.022
Free sulfur dioxide	2 ~ 289	132	35.308	17.007
Total sulfur dioxide	9 ~ 440	251	138.361	42.498
Density	0.987 ~ 1.039	890	0.994	0.003
pH	2.72 ~ 3.82	103	3.188	0.151
Sulphate	0.22 ~ 1.08	79	0.49	0.114
Alcohol	8 ~ 14.2	103	10.514	1.231
Quality	11 nominal values (0 ~ 10)			

Table 10. The number of functional dependencies for each attribute on the right-hand side of the white wine data set

RHS attribute of FDs	The number of FDs having adjusted R-squared value of less than 20%	The number of FDs having adjusted R-squared value of 20% ~ 50%	The number of FDs having adjusted R-squared value of 50% ~ 82%	Total & explanatory FD ratio
Density	2	12	49	63 (96.8%)
Citric_acid	34	0	0	34 (0%)
pH	44	19	1	64 (31.3%)
Free_sulfur_dioxide	14	41	0	55 (74.5%)
Volatile_acidity	35	0	0	35 (0%)
Chloride	47	0	0	47 (0%)
Sulphate	92	0	0	92 (0%)
TOTAL	268	72	50	390 (30.7%)

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The author contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The author has no conflict of interest to declare.