Comparing Classification Algorithms

Zeynep Karkıner
Computer Engineering Department, Baskent University
Ankara, Turkey
21995712@mail.baskent.edu.tr

Abstract— **Improving the wine quality and categorize wine types according to their physical and chemical features are crucial for the wine industry. In this paper, several classification algorithms are used in the wine dataset in order to classify them into three classes. The evaluation process is handled with the use of accuracy, f1-score, precision and recall metrics. As mentioned in the discussion section, the usage of feature selection and cross validation techniques are important factors due to achieving better results.**

Keywords—multiclass-classification, classification algorithms, supervised learning, wine classification

I. Introduction

Classification problems are can be considered as most common research area of the machine learning field. Classification problems are divided into several subtopics such as binary classification, multiclass classification etc. Classification is a branch of supervised learning i.e.; model will be training with the actual values of that can be predicted in the future. Classification algorithms can be used in many fields such as, to categorize application is harmful or not, examining tweets which can be positive or offensive, and categorize wine types which is subject of this paper.

The purpose of this paper is to compare classification algorithms in the wine dataset. The comparison will be evaluated via metrics that are accuracy, precision, f1- score, and recall. The wine dataset was gathered from UCI Machine Learning Repository [1]. This dataset is created on chemical analysis of three different wine cultivars grown in Italy. The wine dataset has 178 instances which can be considered as a small dataset. The wine dataset includes following attributes:

1. Alcohol
2. Malic acid
3. Ash
4. Alkalinity of ash
5. Magnesium
6. Total phenols
7. Flavonoids
8. Nonflavonoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

All attributes are continuous, and the class attribute identifies the class of wine which can be either 0, 1, or 2. This is illustrating the classification type which is multiclass classification. In multiclass classification there will be more than one classes.

For the classification task, Logistic Regression, K-Nearest Neighbor (KNN), Naive Bayes, and Random Forest algorithms are used.

While executing the experiments in KNN and random forest algorithm, some additional experiments have been done for the more promising results. For instance, in KNN experiment, the number of neighbors parameter's initial value was set on three and this value is incremented until accuracy incrementation is stopped. A similar approach is used in the random forest algorithm, the only difference is the subject of the experiment is changed with number of estimators parameter.

This paper is consisting of four parts. According to the outline, introduction section covers the classification problem and dataset information. Literature reviews are presented in second section. In order to understand the usage of classification algorithms, some papers reviewed in the research field. In this section, various research papers are examined due to usage of classification algorithms. UCI repository has provided some papers that used wine dataset. After that, detailed explanation of paper and comparison evaluations are stated in Finding section. Each of the algorithm results are presented in its own table which includes evaluation metrics. Lastly, conclusion and discussions are presented in fourth section. This section includes the summarization of the paper and two discussion topics. Following section is the references that will used while preparing the paper.

II. Literature Review

Gorad Sudhir et. al, proposed a paper that compares classification algorithms specifically, decision trees, K-Nearest Neighbor, Naïve Bayes, Neural Networks and Support Vector Machine (SVM). Comparison is based on their advantages and disadvantages [2].

Rahman R et. al, proposed a study that the detection of permission-based malware with using machine learning techniques. They used Naïve Bayes, KNN, Decision Tree, Random Forest and Decision Forest methods and combined with features picked from the retrieval of Android permissions to categorize applications which is harmful or not [3].

Claster, W. B., Caughron, M., & Sallis, P. J., studied a sentiment analysis according to the tweets which is about red wine. Their purpose is, corroborate industry sales figures with using artificial neural networks technique [4].

Shruthi P., proposed a comparison paper that is to examine wine quality with using classification techniques namely, Naïve Bayes, Simple Logistic, KStar, JRip and J48. Shruthi P., classified wine quality into three main categories and compares the algorithms [5].

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©2022 IEEE

Liu Y., applied a gradient boosted model to make predictions on wine dataset which includes both red and white wine samples. Author states, this study as a multiclass classification but, gradient boost models can be applied on regression models too. By using this model, high accuracy scores are achieved in the no error allowed situation [6].

Aich S. et. al, proposed a new approach by considering feature selection algorithms respectively, Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE). Analyzing the performance metrics are handled with the use of nonlinear decision tree-based classifiers. Their aim is, to help the wine experts to understand the importance factors of while selecting the good quality wine [7].

III. Findings

This section covers a route that starts with data preprocessing and continues with evaluations of each classification algorithm mentioned in the Introduction section.

Firstly, to ease the preprocessing, the extension of wine.data is changed to the wine.csv. This process will increase the understandability and readability of dataset. Attribute names are expressed in integers in original dataset. To make clearer, columns can be changed into real column names mentioned in the Introduction section. Optionally, all preprocessing process can be skipped with the use of Sklearn *datasets* library. It includes lots of datasets which contains the wine dataset. Data preprocessing process can be finished with the check of existing any missing values. This dataset does not contain any missing values.

Examine the correlation between attributes express lots of information. Visualization of the correlation between attributes can be done via a heatmap. A heatmap of the wine dataset is shown in Figure 1.

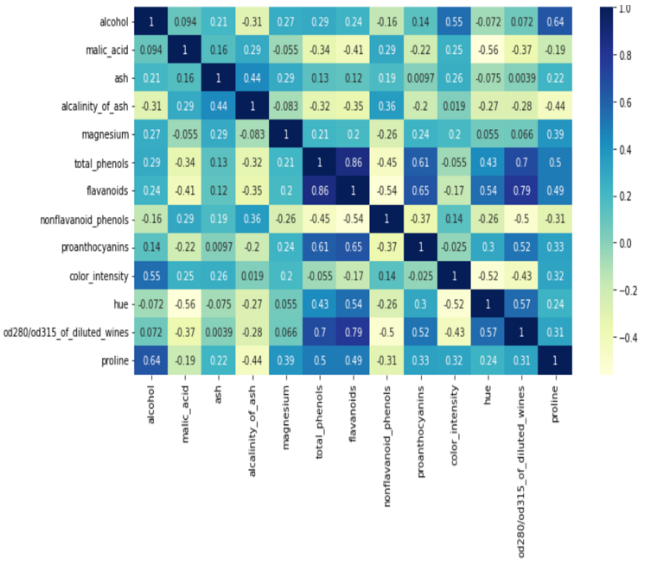

FIGURE 1. HEATMAP OF THE WINE DATASET

According to the correlation matrix and visualized in the heatmap, Ash attribute is the least relevant attribute while comparing the other attributes. Handling the feature selection part will be discussed in Conclusion and Discussion Section.

After this process, the target is the determine class types which are respectively 0,1, and 2. Each class type identifies a kind of wine. The use of two data frames will be a solution for this problem. First data frame contains target classes and the second data frame is consisting of rest of the dataset. The integrity of the dataset is created by doing the concatenation operation of these two data frames.

Before starting the train-test split process, there should be a data frame (X) which consist of the rather than class type attributes. This data frame should contain attributes such as alcohol, malic acid etc. On the other hand, there must be a data frame includes only class types i.e., 0,1, and 2. These are the target values that the classification algorithms will predict. (Y)

The phase of dividing the dataset into two parts which are train data and test data. The implementation has done with the use of Sklearn *model selection* library. Test size has chosen as 0.2 which is quite reasonable ratio for under the consideration of working with a small dataset. At the end of this phase, fitting of the model process may begin.

For model fitting process, train data is used. On the other hand, for the prediction process test data is used. This operation is applied on each classification algorithm. Firstly, logistic regression algorithm is applied on wine dataset. logistic regression is a process of modeling the probability of a discrete outcome given an input variable [8]. Logistic regression generally used for the binary classification tasks but it can be implemented in multiclass classification tasks. The results are that belongs to logistic regression algorithm has shown in Table1. According to the Table 1, the usage of logistic regression produced a high accuracy score.

TABLE 1. LOGISTIC REGRESSION ALGORITHM EVALUATION

	Evaluation Metrics			
	Class	***Precision***	***Recall***	***F1-score***
Logistic Regression	0	1.00	1.00	1.00
	1	0.93	1.00	0.97
	2	1.00	0.90	0.95
Accuracy	0.97222			

Secondly, the study case is the use of K-Nearest Neighbor algorithm. KNN, uses the proximity to make classification or predictions about the grouping of an individual data point [9].

C.C, used various version of KNN classifiers to make predictions in heart disease dataset. The KNN was the best with an accuracy of 69% [10].

Yunneng, Q. proposed an improved KNN algorithm which is used in the stock price prediction. Yunneng states, traditional KNN uses the latest day data to predict the change trend of the next day. However, improved KNN shares the price information group of the first N days is synthesized into a sample which is an input to the KNN model for learning [11].

In this study, KNN algorithm tested with several *n_neighbor* numbers which are respectively; 3, 5 and 9. The results have presented in Table [2-4].

TABLE 2. KNN (n=3) ALGORITHM EVALUATION

	Evaluation Metrics			
	Class	*Precision*	*Recall*	*F1-score*
KNN ***(n_neighbors=3)***	0	0.86	1.00	0.92
	1	0.75	1.00	0.69
	2	0.60	0.90	0.60
Accuracy	0.75			

TABLE 3. KNN (n=5) ALGORITHM EVALUATION

	Evaluation Metrics			
	Class	*Precision*	*Recall*	*F1-score*
KNN ***(n_neighbors=5)***	0	1.00	1.00	1.00
	1	0.77	0.71	0.74
	2	0.64	0.70	0.67
Accuracy	0.80555			

TABLE 4. KNN (n=9) ALGORITHM EVALUATION

	Evaluation Metrics			
	Class	*Precision*	*Recall*	*F1-score*
KNN ***(n_neighbors=9)***	0	0.92	1.00	0.96
	1	1.00	0.64	0.78
	2	0.64	0.90	0.75
Accuracy	0.83333			

After this point, incrementation of accuracy value is stopped. Furthermore, it started decrease. Next classification algorithm that will be evaluate is Naive Bayes. Naïve Bayes based on probability models that intercorporate strong independence assumptions these assumptions do not have an impact on reality. Thus, they are considered as *naive* [12].

Kamila, V. Z., Subastian, E., & Rosmasari, studied an optional advanced course recommendation with using KNN and Naïve Bayes algorithms. Authors mentioned about the small dataset which is provided from Mulawarman University. They achieved 100% accuracy in KNN with value of K=1 and, 100% accuracy for the naïve bayes approach [13].

Implementation of Naive Bayes has done via the use of *Gaussian NB* library and also the model with the same name is used. In Table 5, detailed evaluation is presented.

TABLE 5. NAIVE BAYES ALGORITHM EVALUATION

	Evaluation Metrics			
	Class	*Precision*	*Recall*	*F1-score*
Naïve Bayes	0	0.92	1.00	0.96
	1	1.00	0.93	0.96
	2	1.00	1.00	1.00
Accuracy	0.97222			

Naive Bayes and Logistic Regression algorithms produced the same accuracy score. However, they are differing with their precision, recall and F1- score metrics.

Final classification algorithm that will be discussed is the Random Forest algorithm. Random Forest combines the output of multiple decision trees to reach a single result. Random Forest classifier model takes a parameter which is *n_estimator*. This is the number of trees that build before taking the maximum voting.

Reddy, P. D. and Parvathy, L. R., studied research which is a prediction analysis using Random Forest algorithms to forecast the air pollution level in particular location. Study is based on the comparison between Random Forest and Naive Bayes algorithms. In conclusion, they achieved random forest performed better than Naive Bayes under the forecasting of air pollution level [14].

In this classification task, initial number of estimators value is setting as two, presented in Table 6.

TABLE 6. RANDOM FOREST (n=2) ALGORITHM EVALUATION

	Evaluation Metrics			
	Class	*Precision*	*Recall*	*F1-score*
Random Forest ***(n_estimators = 2)***	0	0.80	1.00	0.89
	1	1.85	0.79	0.81
	2	1.00	0.80	0.89
Accuracy	0.86111			

Incrementation of the value of n, will bring higher accuracy score at one point. For n=6, maximum accuracy score is achieved shown in the Table 7.

TABLE 7. RANDOM FOREST (n=3) ALGORITHM EVALUATION

	Evaluation Metrics			
	Class	*Precision*	*Recall*	*F1-score*
Random Forest ***(n_estimators = 6)***	0	0.86	1.00	0.92
	1	1.00	0.86	0.92
	2	1.00	1.00	1.00
Accuracy	0.94444			

Up to now, four classification algorithms evaluated with different parameters and metrics. A summary of four classification algorithms have presented in Table 8.

TABLE 8. SUMMARY OF CLASSIFICATION ALGORITHMS

Classification Algorithm	**Accuracy**
Logistic Regression	0.97222
KNN (k=9)	0.83333
Naïve Bayes	0.97222
Random Forest (n=6)	0.94444

For more promising results, cross validation technique can be used. Cross validation with 10 K-folds is the most common usage in the literature. However, it can be differing due to the specific usage. The main idea behind the cross validation is shuffling the dataset and randomly picks the training and test data in the chosen folds. In this study, splitting process has done with 10 K-Folds. Four classification algorithms respectively, logistic regression, KNN, naïve bayes and random forest are executed with the use of cross validation. In Table 9, accuracy scores with the use of cross validation technique denoted.

TABLE 9. ACCURACY SCORES WITH THE USE OF CROSS VALIDATION

Classification Algorithm	Accuracy
Logistic Regression	0.95323
KNN (k=3)	0.71089
Naïve Bayes	0.96858
Random Forest (n=7)	0.97564

Under the considerations of these conditions; random forest algorithm gave higher accuracy scores with the use of cross validation. All comparing process is done with the produced highest accuracy score of each algorithm.

IV. CONCLUSIONS AND DISCUSSIONS

In conclusion, this paper is based on comparing the classification algorithms in the multiclass classification task. The task is to predict the wine's class type which can be either 0,1 or 2. Wine dataset does not contain any missing value this can be interpreted as require less effort on the data preprocessing process. There are several classification algorithms tested in wine dataset and all scores that achieved are presented in the tables. According to the findings, without the cross validation, logistic regression and naïve bayes algorithms produced the highest accuracy score which are both 97% in the wine dataset. However, random forest and naïve bayes algorithms produced a high score with the use of cross validation which are respectively, 97% and 96%.

To summarize all paper, 4 classification algorithms which were split with and without cross validation and their accuracy scores denoted in the Table 10.

TABLE 10. SUMMARY OF THE STUDY

Classification Algorithm	Accuracy (w/Cross Validation)	Accuracy (w/o Cross Validation)
Logistic Regression	0.95323	0.97222
KNN	0.71089	0.83333
Naïve Bayes	0.96858	0.97222
Random Forest	0.97564	0.94444

Results can be optimized with two ways which will create the discussion part of the section.

First discussion topic is the use of Ash attribute. Correlation matrix calculates each of the correlations between attributes. According to the findings, Ash attribute is the least relevant with the dataset. Thus, eliminating this attribute may increase the value of accuracy in each classification algorithms.

Second discussion topic is the use of cross validation technique on the while separating the dataset into two parts which are training and test data. Wine dataset has small numbers of attributes and instances as mentioned in the several sections. Thus, overfitting may occur very fast. Cross validation technique prevents the overfitting issue with according to the number of folds.

REFERENCES

[1] Forina, M.,1991-07-01, "Wine Data Set," Institute of Pharmaceutical and Food Analysis and Technologies, Via Brigata Salerno, 16147 Genoa, Italy. Available: https://archive.ics.uci.edu/ml/datasets/wine

[2] Gorade S., "A Study Some Data Mining Classification Techniques". (2017). *International Journal of Modern Trends in Engineering & Research*, 210–215. https://doi.org/10.21884/ijmter.2017.4031.zt9tv

[3] Rahman, R., Islam, M. R., Ahmed, A., Hasan, M. K., & Mahmud, H. (2022). A study of permission-based malware detection using machine learning. *2022 15th International Conference on Security of Information and Networks (SIN)*. https://doi.org/10.1109/sin56466.2022.9970528

[4] Claster, W. B., Caughron, M., & Sallis, P. J. (2010). Harvesting consumer opinion and wine knowledge off the social media grape vine utilizing artificial neural networks. *2010 Fourth UKSim European Symposium on Computer Modeling and Simulation*. https://doi.org/10.1109/ems.2010.109

[5] Shruthi, P. (2019). Wine quality prediction using data mining. *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*. https://doi.org/10.1109/icatiece45860.2019.9063846

[6] Liu, Y. (2021). Optimization of gradient boosting model for wine quality evaluation. *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*. https://doi.org/10.1109/mlbdbi54094.2021.00033

[7] Aich, S., Al-Absi, A. A., Hui, K. L., Lee, J. T., & Sain, M. (2018). A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques. *2018 20th International Conference on Advanced Communication Technology (ICACT)*. https://doi.org/10.23919/icact.2018.8323673

[8] *Logistic regression*. | ScienceDirect Topics. (n.d.). Retrieved December 18, 2022, from https://www.sciencedirect.com/topics/computer-science/logistic-regression

[9] *K-nearest neighbors algorithm*. IBM. (n.d.). Retrieved December 18, 2022, from https://www.ibm.com/topics/knn

[10] C, C. (2021). Prediction of heart disease using different KNN classifier. *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*. https://doi.org/10.1109/iciccs51141.2021.9432178

[11] Yunneng, Q. (2020). A new stock price prediction model based on improved KNN. *2020 7th International Conference on Information Science and Control Engineering (ICISCE)*. https://doi.org/10.1109/icisce50968.2020.00026

[12] Naive Bayes. Retrieved December 18, 2022, from https://www.ibm.com/docs/en/ias?topic=procedures-naive-bayes

[13] Kamila, V. Z., Subastian, E., & Rosmasari. (2019). KNN and Naive Bayes for optional advanced courses recommendation. *2019 International Conference on Electrical, Electronics and Information Engineering (ICEEIE)* https://doi.org/10.1109/iceeie47180.2019.8981450

[14] Reddy, P. D., & Parvathy, L. R. (2022). Prediction analysis using random forest algorithms to forecast the air pollution level in a particular location. *2022 3rd International Conference on Smart Electronics and Communication (ICOSEC)*. https://doi.org/10.1109/icosec54921.2022.9952138