1. Tsehay Admassu ASSEGIE[1], 2. Sangeetha MURUGAN[2], 3. Rajkumar GOVINDARAJAN[3], 4. Komal Kumar NAPA[4], 5. Nageswari D[5]

School of Electronic and Electrical Engineering, Kyungpook National University, Daegu, Republic of Korea (1)
Department of Computer Science & Engineering, Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India (2)
Department of Computer Science & Engineering (Data Science), Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India (3)
Department of Computer Science & Engineering (Data Science), Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India (4)
Department of Science & Humanities (General Engineering-EEE), R.M.K. College of Engineering and Technology, Thiruvallur, Tamil Nadu, India (5)
ORCID: 1. https://orcid.org/0000-0003-1566-0901; 4. https://orcid.org/0000-0001-8662-0224

doi:10.15199/48.2024.07.29

Improving the Performance of Machine Learning with Sequential Feature Selection and Grid Search

***Abstract**. Feature selection is an important step in developing accurate machine-learning models for classification tasks, including wine quality prediction. The accuracy of the machine learning model depends on the selection of relevant features that contribute to the predicted outcome. In this paper, we propose two commonly used optimization methods, forward sequential feature selection (SFS), and grid search, to identify the most relevant features for wine quality prediction using K-nearest neighbor (KNN). We used a dataset of 1598 samples with 11 wine-quality features and evaluated the performance of the KNN model trained on different subsets of features selected SFS. The result suggests that SFS and gird search are effective methods for wine quality prediction using KNN. The identified wine quality features help to predict the quality of wine more accurately, leading to better predictive outcomes. Thus, machine learning models can benefit greatly from the use of grid search and SFS. By fine-tuning the model in this way, it is possible to achieve better results in applications where accuracy and speed are important. As machine learning continues to be used in a wide range of industries, the use of these techniques will become increasingly important. Further research is needed to validate the model on larger datasets and to integrate it into practical classification or predictive analysis.*

Streszczenie.** Wybór funkcji to ważny krok w opracowywaniu dokładnych modeli uczenia maszynowego do celów klasyfikacji, w tym przewidywania jakości wina. Dokładność modelu uczenia maszynowego zależy od wyboru odpowiednich cech, które przyczyniają się do przewidywanego wyniku. W tym artykule proponujemy dwie powszechnie stosowane metody optymalizacji, sekwencyjny wybór cech w przód (SFS) i przeszukiwanie siatki, aby zidentyfikować cechy najbardziej odpowiednie do przewidywania jakości wina za pomocą K-najbliższego sąsiada (KNN). Wykorzystaliśmy zbiór danych obejmujący 178 próbek z 13 cechami jakości wina i oceniliśmy działanie modelu KNN wyszkolonego na różnych podzbiorach wybranych cech FSFS. Wynik sugeruje, że SFS i przeszukiwanie pasów są skutecznymi metodami przewidywania jakości wina za pomocą KNN. Zidentyfikowane cechy jakości wina pomagają dokładniej przewidzieć jakość wina, co prowadzi do lepszych wyników predykcyjnych. Zatem modele uczenia maszynowego mogą w znacznym stopniu skorzystać na wykorzystaniu wyszukiwania siatki i SFS. Dostrajając w ten sposób model, możliwe jest osiągnięcie lepszych wyników w zastosowaniach, w których ważna jest dokładność i szybkość. Ponieważ uczenie maszynowe jest w dalszym ciągu wykorzystywane w wielu gałęziach przemysłu, wykorzystanie tych technik będzie zyskiwać na znaczeniu. Konieczne są dalsze badania, aby zweryfikować model na większych zbiorach danych i włączyć go do praktycznej klasyfikacji lub analizy predykcyjnej.* (Poprawa wydajności uczenia maszynowego dzięki sekwencyjnemu wyborowi funkcji i przeszukiwaniu siatki***)

Keywords: K-Nearest Neighbors, parameter tuning, machine learning.
Słowa kluczowe: akość wina, dostrajanie parametrów, uczenie maszynowe.

Introduction

In the problem of wine quality prediction, using grid search and sequential feature selection can lead to more accurate predictions and a better understanding of the factors that contribute to wine quality [1]. By optimizing hyperparameters and selecting the most relevant features, the model can better understand the relationships between the various chemical and physical properties of wine and its perceived quality.

For instance, grid search improves the parameters of a support vector machine (SVM) model for wine quality prediction can lead to significant improvements in accuracy. A study by S.F. Radz et al. [2] found that using grid search to optimize the C and gamma parameters of an SVM model led to an increase in accuracy by 12% for breast cancer prediction.

Similarly, using sequential feature selection can help to identify the most important features for predicting wine quality. A study by S.S Subbiah et al. [3] found that using sequential forward selection to select the most informative feature for various machine learning classifiers (such as extreme boosting, support vector machine, KNN, and random forest led to an increase in accuracy from 86.3% to 98% using an extreme boosting model for chronic heart disease prediction.

Furthermore, the study conducted by S. Zaza et al. [4] reported that the feature selection is significant in applying analytical tests to wine quality. However, feature selection methods such as SFS involve determining the feature subset iteratively to find the optimal feature subset. The iterative process of finding the optimal feature subset consumes more computational time.

To address this gap, we propose a method to enhance the performance of machine learning models by combining SFS and grid search techniques. Sequential feature selection does not lead to the optimal subset of features, and SFS can be used to iteratively select features that improve the model's performance. Grid search is then used to find the best hyperparameters for the model. Overall this study is aimed at combining feature selection with a grid search approach to improve the performance of the KNN model for wine quality prediction.

The contribution of this work is as follows: we have trained the KNN classifier and compared its performance on a wine dataset, then we move further to use the SFS method to select the optimal feature subset in the wine dataset and then retrain the model. SFS improved the accuracy of the KNN classifier resulting from 88.5% without SFS to 90.15% with SFS, thereby outperforming the KNN classifier.

Related work

While knowledge-based systems rely on human expertise to create the knowledge base, machine learning algorithms can be used to improve their performance. One approach is to use hyperparameter optimization with grid search, which involves identifying the optimal parameter from a set of parameters that is most relevant to the problem at hand. This can help reduce the complexity of the knowledge base and improve the accuracy of the system [5].

In the context of wine quality prediction, grid search is used to optimize the parameters of the system [6]. Grid search involves testing different combinations of parameter values to find the combination that results in the best performance. This can help fine-tune the system and improve its ability to make accurate decisions and solve problems

Sequential feature selection is a technique used to select the most important features for a given task [7, 8]. In the context of wine quality prediction, this technique can be used to identify the most important wine characteristics for predicting quality.

For example, a study conducted by Y. Gupta [9] used correlational feature selection to identify the most important wine characteristics for predicting wine quality. The study found that acidity levels, alcohol percentage, and sulfates were the most important features for predicting wine quality.

Moreover, grid search is a technique used to optimize the parameters of a machine learning algorithm [10]. In the context of wine quality prediction, grid search can be used to find the optimal configuration of a machine learning algorithm that maximizes its accuracy [11].

By using grid search, the performance of the machine learning algorithm can be improved, which can lead to more accurate predictions of wine quality. For example, a study conducted by [12] used grid search to optimize the parameters of a support vector machine algorithm for wine quality prediction. The study found that the optimal configuration of the algorithm was achieved with a linear kernel and a C value of 1.

Previous studies have focused on either feature selection or grid search individually for wine quality prediction. While both techniques have shown promising results, there is a research gap in combining these two approaches. The uncombined feature selection and grid search methods may not be able to find the optimal combination of features and algorithm parameters, leading to suboptimal performance. Therefore, there is a need for research that explores the potential benefits of combining these two techniques for wine quality prediction.

Hybrid grid search and sequential feature selection can further improve the performance of machine learning algorithms for wine quality prediction. By combining these two techniques, the most important features can be selected and the optimal configuration of the algorithm can be found simultaneously. This approach can lead to even more accurate predictions of wine quality, as it takes into account both the relevance of the features and the best parameters for the algorithm. This hybrid approach can also reduce overfitting, as it selects only the most relevant features and optimizes the algorithm's parameters accordingly. In conclusion, a hybrid approach of grid search and sequential feature selection can improve the accuracy and efficiency of machine learning algorithms for wine quality prediction, providing valuable insights for winemakers and consumers.

Methodology

The proposed hybrid feature selection and grid search-based wine quality prediction model involves five steps. Firstly, the wine quality data is collected from publicly available sources (Kaggle) repository and pre-processed to remove missing values, outliers, and redundant features. Secondly, various feature selection techniques such as correlation-based feature selection, and recursive feature elimination will be applied to identify the most relevant features for wine quality prediction. Thirdly, a grid search approach is used to explore the optimal combination of hyperparameters for different machine learning algorithms such as support vector machines, random forests, and neural networks. Fourthly, the performance of the wine quality prediction model is evaluated using various metrics such as accuracy, precision, recall, and F1 score. The models will also be compared with existing wine quality prediction models to assess their performance. Finally, the results obtained from the experiments are analyzed to identify the best-performing model and the most relevant features for wine quality prediction.

The wine quality dataset contains 11 features and 1598 samples. Table 1 indicates the features used in developing the classification model.

Table 1. The description of the wine quality dataset feature

No.	Feature	Description
1.	Fixed acidity	Acid is not reality evaporable
2.	Acidity	The amount of acetic acid in wine
3.	Citric acid	Citric acid in wine
4.	Residual sugar	The amount of sugar after fermentation
5.	Chloride	The amount of salt in wine
6.	Sulfur dioxide	The free form of SO2
7.	Total sulphur dioxide	Free and bound forms of SO2
8.	Density	The density of water
9.	PH	Indicate the acidity or basicity of wine
10.	Sulfates	Wine additive which acts as antimicrobial
11.	Free sulfur dioxide	The amount of sulfur dioxide (SO2) in free form

To get further insight into the wine quality dataset, a statistical analysis such as Pearson correlation was employed to analyze the relationship between the dataset features. The Pearson correlation is important because the SFS may select redundant features and this can affect its effectiveness when dealing with highly correlated features. The Pearson correlation is determined by the formula defined in Equation 1.

$$r=\frac{\sum\left(Xi-\overline{Xi}\right)\left(Y_i-\overline{Y_i}\right)}{\sqrt{\sum_{i=1}^{N}\left(Xi-\overline{X_i}\right)^2+\left(Y_i-\overline{Y_i}\right)^2}} \quad (1)$$

Where: *N*– number of wine quality features, *r* – correlation coefficients, X-values of the independent variable X, Y-values, $\overline{X}$ - means of variable X , and $\overline{Y}$ - mean of the values of the variable y in the wine quality dataset.

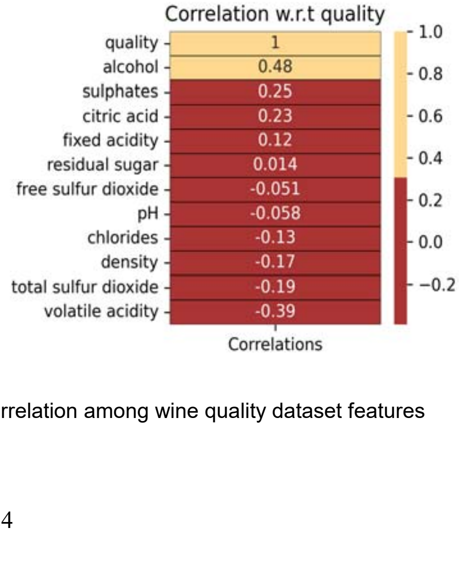

Fig.1. Correlation among wine quality dataset features

Figure 1 presents the correlation among wine quality datasets. As presented in Figure 1, features such as alcohol, sulfates, critic acid, fixed acidity, and residual sugar have a strong relationship with the wine quality. In discriminately, pH, free sulfur dioxide, density, chlorides, total sulfur dioxide, and volatile acidity have a negative relationship with the wine quality.

Figure 2 presents the relationship between the independent and the dependent feature (the wine quality). From the visualization of the bar plot of the relationship between the dependent and the independent feature of wine quality, it can be observed that the quality increases with an increase in alcohol, citric acid, and sulfates. However, the wine quality decreases as the features such as volatile acidity, chlorides, and pH increase. Furthermore, free sulfur dioxide and total sulfur dioxide alone will not predict the quality of wine as there is no clear pattern on their increase and decrease and the quality of the wine.

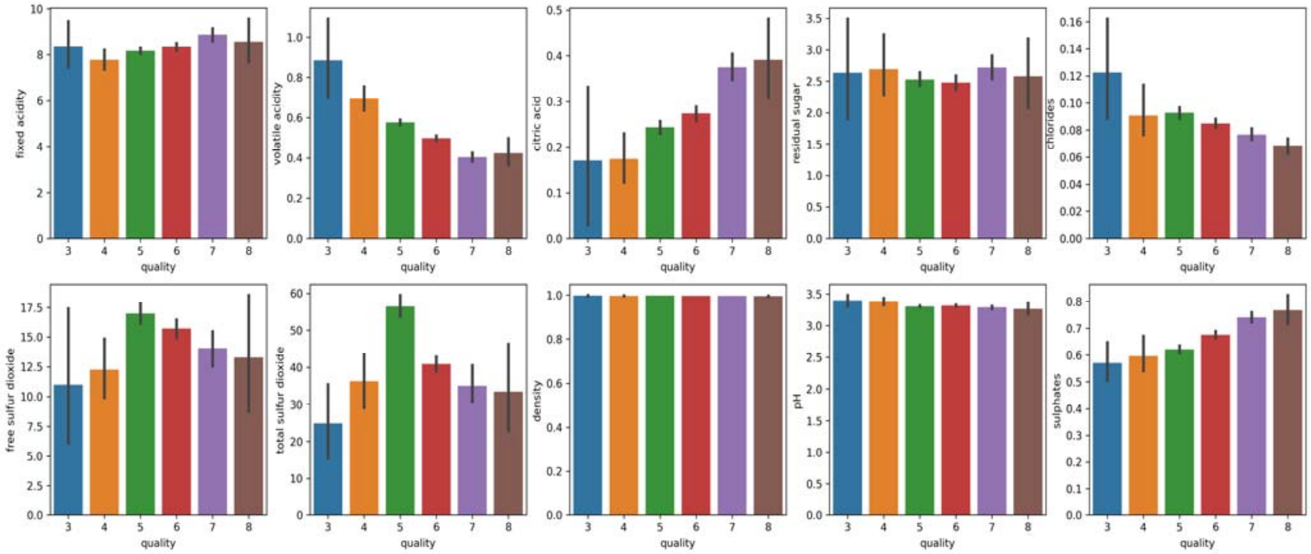

Fig.2. Validation accuracy vs number of features

Result and discussion

This section presents the experimental results of the proposed SFS and Grid Search-based approach to enhance the performance of machine learning models by combining SFS and grid search techniques. The authors argue that SFS leads to the optimal subset of features and that SFS can be used to iteratively select features that improve the model's performance. Grid search is then used to find the best hyperparameters for the model. The use of both SFS and grid search technique outperforms compared In the specific case of wine quality prediction, using grid search and SFS leads to more accurate predictions and a better understanding of the factors that contribute to wine quality. By optimizing hyperparameters and selecting the most relevant features, the model can better capture the complex relationships between the various chemical and physical properties of wine and its perceived quality.

The combination of grid search and SFS can be used together to improve the performance of machine learning models such as the KNN classifier. By fine-tuning the hyperparameters and selecting the most relevant features, the KNN model can be optimized for better accuracy and efficiency to the individual feature selection or grid search technique for parameter tuning.

The baseline KNN performed with an accuracy of 88.54%, and an ROC score of 76.58% using the distance metric and K=2. With hyperparameter tuning the KNN achieved an accuracy score of 91.66% using K=2, and K-weights of distance metric.

Conclusion

In conclusion, the paper suggests that combining SFS and grid search improves the performance of machine learning models, such as the KNN classifier for wine quality prediction. The proposed method outperforms traditional feature selection methods and can be used as an effective tool for improving the accuracy of machine learning models. The authors argue that traditional feature selection methods may not always lead to the optimal subset of features and that SFS can be used to iteratively select features that improve the model's performance with grid search to find the best hyperparameters for the model.

The proposed method is tested on the wine quality dataset and using the KNN classifier. The results show that the proposed method outperforms traditional feature selection methods and improves the accuracy of the KNN in identifying the wine as bad or good based on the features. Overall, the paper suggests that combining SFS and grid search can lead to better performance in machine learning models. Sequential feature selection can help identify the most important features for predicting wine quality, such as acidity levels, sugar content, and alcohol percentage.

It is conclusive that the SFS and grid search can be used to improve the performance of machine learning algorithms for wine quality prediction. Sequential feature selection can help identify the most important wine characteristics for predicting quality, while grid search can be used to find the optimal configuration of a machine learning algorithm that maximizes its accuracy.

By combining knowledge-based and machine-learning approaches, more accurate predictions of wine quality can be achieved, which can benefit both winemakers and consumers.

***Authors**:*

Mr. Tsehay Admassu ASSEGIE, School of Electronic and Electrical Engineering, Kyungpook National University, Daegu, Republic of Korea. *Corresponding Author E-mail: tsehayadmassu2006@gmail.com.*

Rajkumar GOVINDARAJAN, Department of Computer Science & Engineering (Data Science), Madanapalle, Andhra Pradesh, India.

Sangeetha MURUGAN, Department of Computer Science & Engineering, Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India.

Komal Kumar NAPA, Department of Computer Science & Engineering (Data Science), Madanapalle Institute of Technology & Science, Madanapalle, Andhra Pradesh, India
Nageswari D., Department of Science & Humanities (General Engineering-EEE), R.M.K. College of Engineering and Technology, Thiruvallur, Tamil Nadu, India

REFERENCES

[1] Dahal R., Dahal N. Banjade H. Gaire S., Prediction of Wine Quality Using Machine Learning Algorithms, *Open Journal of Statistics,* 2021, 11, 278-289 https://www.scirp.org/journal/ojs

[2] Radzi S. et al.,, Hyperparameter Tuning and Pipeline Optimization via Grid Search Method and Tree-Based AutoML in Breast Cancer Prediction, J. *Pers. Med*. 2021, 11, 978. https://doi.org/10.3390/ jpm11100978.

[3] Siva S., Jayakumar C., Opportunities and Challenges of Feature Selection Methods for High Dimensional Data: A Review, *Ingénierie des Systèmes d'Information*, vol. 26, No. 1, February, 2021, pp. 67-77.

[4] Siphendulwe Z., Marcellin A., Sisipho H., Wine feature importance and quality prediction: A comparative study of machine learning algorithms with unbalanced data, arXiv: 2310.01584v1 [stat.AP] 2 Oct 2023.

[5] Yasser A., Emad A., Muna A., and Ali M., Hyperparameter Search for Machine Learning Algorithms for Optimizing the Computational Complexity, *Processes*. 2023, 11, 349. https:// doi.org/10.3390/pr11020349.

[6] Khushboo J., Keshav K., Sachin K., Shubham M., Seifedine K., Machine learning-based predictive modeling for the enhancement of wine quality, *Scientific Reports*, 2023, https://doi.org/10.1038/s41598-023-44111-9.

[7] Jörn L., and Alfred U., Enhancing Explainable Machine Learning by Reconsidering Initially Unselected Items in Feature Selection for Classification, *Bioinformatics,* 2022, 2, 701–714. https://doi.org/10.3390/biomedinformatics2040047.

[8] Ritu A., and Saurabh P., Sequential Feature Selection and Machine Learning Algorithm-Based Patient's Death Events Prediction and Diagnosis in Heart Disease, *SN Computer Science*, 2020, https://doi.org/10.1007/s42979-020-00370-1.

[9] Yogesh G., Selection of Important and Predicting Wine Quality Using Machine Learning Techniques, *Porceedia Computer Science*, 2018, 10.1016/j.procs.2017.12.041.

[10] Terry C., Chienwen W., and Chun C., A Generalized Wine Quality Prediction Framework by Evolutionary Algorithms, *International Journal of Interactive Multimedia and Artificial Intelligence*, Vol. 6, April 2021.

[11] Wiharto, Esti S., and Sigit S., Framework Two-Tier Feature Selection on the Intelligence System Model for Detecting Coronary Heart Disease, *Ingénierie des Systèmes d'Information*, vol. 26, No. 6, December, 2021, pp. 541-547.

[12] Sathishkumar M., Reshmy K., Sabaria S., Prasannavenkatesan T., An investigation of wine quality testing using machine learning techniques, *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, June 2023, pp. 747~754 ISSN: 2252-8938, DOI: 10.11591/ijai.v12.i2.pp747-754.