ORDINAL REGRESSION WITH A TABULAR WINE QUALITY MODELS REPORT

Name	ID	Name	ID
Zayyam Fida	21032133	Ahtsham Karim	21031263
Hamza Raqeeb	21031287	Ali Akbar	21034385
Adeel Hussain Shah	21031305	Aqib Khurshid	21016715

Dataset

The dataset used in the contest, consisting of both training and test data, was obtained from a deep learning model that had been previously trained on the Wine Quality dataset. While the features of the contest dataset show similarity to the original dataset, they are not entirely identical. Participants in the competition are strongly encouraged to make use of the original dataset to examine the variations and evaluate how incorporating it into the training process affects the performance of their models. [1].

Python libraries

Python libraries are essential for expanding the capabilities of the Python programming language, enabling efficient development and access to pre-built functions and tools for various tasks. They promote code reuse, accelerate development cycles, and empower developers to easily tackle complex problems [2].

Train Qualities Count

The term "train qualities count" refers to the overall assessment of a train's desirable attributes and characteristics. These qualities encompass factors such as speed, reliability, comfort, safety features, efficiency, and capacity. A higher count denotes a superior level of train quality and performance.

Full train shape	**(3199, 12)**

Statistical Description

Statistics provide valuable insights about the distribution and characteristics of data in a training dataset, helping to identify patterns, outliers, and potential biases, which is crucial for making informed decisions during the training process and improving the accuracy and reliability of machine learning models[3]. Figure 1: Shown Statistical description

```
full_train.describe().iloc[1:,:-1]
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
mean	8.345858	0.528936	0.266239	2.446499	0.08367	16.477024	48.049859	0.996742	3.310728	0.647168	10.424669
std	1.720511	0.175484	0.191294	1.065161	0.03414	10.115222	32.930681	0.001863	0.147582	0.150524	1.048119
min	4.600000	0.120000	0.000000	0.900000	0.01200	1.000000	6.000000	0.990070	2.740000	0.330000	8.400000
25%	7.100000	0.390000	0.090000	1.900000	0.07100	7.000000	22.000000	0.995580	3.200000	0.550000	9.500000
50%	7.900000	0.520000	0.250000	2.200000	0.07900	15.000000	42.000000	0.996700	3.310000	0.620000	10.100000
75%	9.100000	0.640000	0.420000	2.600000	0.09000	23.000000	64.000000	0.997800	3.390000	0.720000	11.100000
max	15.900000	1.580000	1.000000	15.500000	0.61100	68.000000	289.000000	1.003690	4.010000	2.000000	14.900000

Figure 1: Shown Statistical description

Table 1: Show `Training Dataset value`

Full Dataset Training quality	value counts
5	1322
6	1240
7	476
4	88
8	55
3	18

QUALITY OF VALUE IN THE DATASET

The wine quality dataset, a limit above which we will log transform refers to a specific value or threshold in a particular feature of the dataset. If any data point in that feature exceeds this limit, a log transformation can be applied to normalize the distribution and reduce the impact of extreme values, promoting better analysis and modeling of the wine quality data moreover [4]. Fig: 2 shows the skew limit

	Skew
chlorides	6.986836
residual sugar	4.557520
sulphates	2.216731
total sulfur dioxide	1.403604
fixed acidity	0.989993
free sulfur dioxide	0.876612
alcohol	0.818845

Fig: 2 shows the skew limit

MIN-MAX SCALER APPROACH

The min-max scaler approach as a data normalization technique used to rescale the features of a dataset within a specific range, usually between 0 and 1. This process involves subtracting the minimum value of each feature and dividing it by the range. By applying this technique, the aim is to achieve consistent scaling across all features, thereby preserving the relative relationships between data points. Additionally, the min-max scaler helps prevent distortion caused by outliers [5].

Train	(2559, 11)
Test	(640, 11))

DIFFERENT MODELS

Random Forest Classifier

The Random Forest Classifier is a machine-learning technique designed specifically for classification tasks. By harnessing the collective power of multiple decision trees, it effectively predicts outcomes by consolidating their individual outputs. This algorithm capitalizes on the principles of ensemble learning, enabling it to enhance accuracy and effectively tackle intricate data patterns. Such characteristics make the Random Forest Classifier a valuable tool for addressing classification challenges within the realm of machine learning [6].

Model Results

model: RandomForestClassifier ()

	precision	recall	f1-score	support
3	1.00	0.00	0.00	2
4	1.00	0.00	0.00	18
5	0.72	0.77	0.74	271
6	0.57	0.65	0.61	242
7	0.51	0.38	0.43	96
8	1.00	0.09	0.17	11
accuracy			0.63	640
macro avg	0.80	0.31	0.33	640
weighted avg	0.64	0.63	0.61	640

K Neighbors Classifier

The K Neighbors Classifier is a machine learning methodology employed to perform classification tasks. It determines the class of a given data point by examining its closest neighbors within the feature space. By considering the k nearest neighbors, the algorithm assigns the predicted class as the one that occurs most frequently among these neighbors. This approach of leveraging proximity-based relationships in the feature space enables the K Neighbors Classifier to make accurate predictions in classification scenarios. Consequently, this algorithm holds considerable significance in the realm of machine learning research [7].

Model Results

model: KNeighborsClassifier()

	precision	recall	f1-score	support
3	1.00	0.00	0.00	2
4	0.00	0.00	0.00	18
5	0.61	0.69	0.65	271
6	0.47	0.54	0.50	242
7	0.46	0.26	0.33	96
8	1.00	0.00	0.00	11
accuracy			0.53	640
macro avg	0.59	0.25	0.25	640
weighted avg	0.53	0.53	0.51	640

Support Vector Machine

The SVM (Support Vector Machine) Classifier is a machine learning technique specifically designed for classification purposes. It builds a hyperplane within a high-dimensional space to effectively distinguish between various classes of data points. The primary objective of the algorithm is to optimize the margin between this hyperplane and the nearest data points. By doing so, it can adeptly handle classification scenarios with both linear and non-linear decision boundaries. The SVM Classifier's ability to construct such discriminative hyperplanes makes it a powerful tool in the field of machine learning research [8].

Model Results

model: SVC()

	precision	recall	f1-score	support
3	1.00	0.00	0.00	2
4	1.00	0.00	0.00	18
5	0.68	0.75	0.71	271
6	0.52	0.67	0.58	242
7	0.54	0.16	0.24	96
8	1.00	0.00	0.00	11
accuracy			0.59	640
macro avg	0.79	0.26	0.26	640
weighted avg	0.61	0.59	0.56	640

Logistic Regression

Logistic Regression is a statistical modeling approach widely employed in binary classification tasks. It aims to estimate the likelihood of an event transpiring by fitting a logistic function to the input features. By learning the optimal coefficients that effectively distinguish between the two classes, this algorithm exhibits interpretability and finds extensive application across diverse domains. Its capacity to model probabilities and derive meaningful insights makes Logistic Regression a highly regarded technique within the realm of research in machine learning [9].

Model Results

model: LogisticRegression()

	precision	recall	f1-score	support
3	1.00	0.00	0.00	2
4	1.00	0.00	0.00	18
5	0.68	0.78	0.73	271
6	0.53	0.65	0.58	242
7	0.53	0.17	0.25	96
8	1.00	0.00	0.00	11
accuracy			0.60	640
macro avg	0.79	0.27	0.26	640
weighted avg	0.61	0.60	0.56	640

Decision Tree Classifier

The Decision Tree Classifier is a machine learning algorithm specifically utilized for classification tasks. It constructs a model resembling a tree structure by iteratively partitioning the data according to the feature that most effectively distinguishes between the classes. Each internal node in the tree signifies a decision based on a particular feature, while each leaf node corresponds to a class label. This characteristic of organizing information into a hierarchical structure enables the Decision Tree Classifier to provide intuitive insights and facilitate comprehensible decision-making processes. Hence, this algorithm holds significant value in the context of research in machine learning [10].

Model Results

```
model: DecisionTreeClassifier()
```

	precision	recall	f1-score	support
3	0.00	0.00	0.00	2
4	0.05	0.06	0.05	18
5	0.64	0.61	0.62	271
6	0.50	0.50	0.50	242
7	0.41	0.46	0.43	96
8	0.17	0.18	0.17	11
accuracy			0.52	640
macro avg	0.30	0.30	0.30	640
weighted avg	0.53	0.52	0.52	640

Summary:

Based on the evaluation results of the wine quality dataset, the performance of the different models can be summarized as follows:

1. Random Forest Classifier:

- Accuracy: 0.63

- 2. K Neighbors Classifier:

- Accuracy: 0.53

3. Support Vector Machine (SVM) Classifier:

- Accuracy: 0.59

4. Logistic Regression:

- **Accuracy: 0.60**

5. Decision Tree Classifier:

- **Accuracy: 0.52**

Summary

Based on the accuracy scores, the Random Forest Classifier appears to be the best-performing model on the wine quality dataset, followed by Logistic Regression and the SVM Classifier. The K Neighbors Classifier and Decision Tree Classifier have lower accuracy and are less effective in this scenario. In conclusion, the Random Forest Classifier, which combines multiple decision trees through ensemble learning, demonstrates the highest accuracy and is the most suitable model for the wine quality dataset among the models evaluated. It is capable of effectively capturing intricate data patterns and providing accurate predictions in classification tasks.

GOOGLE COLAB LINK:

https://colab.research.google.com/drive/1cFQok24OQnl62qzckzErxlWVUVXVtSdi?usp=sharing

References

1) Dahal, K. R., Dahal, J. N., Banjade, H., & Gaire, S. (2021). Prediction of wine quality using machine learning algorithms. *Open Journal of Statistics*, *11*(2), 278-289.
2) Srinath, K. R. (2017). Python–the fastest growing programming language. *International Research Journal of Engineering and Technology*, *4*(12), 354-357.
3) White, H. (1989). Learning in artificial neural networks: A statistical perspective. *Neural computation*, *1*(4), 425-464.
4) Sick, B., Hathorn, T., & Dürr, O. (2021, January). Deep transformation models: Tackling complex regression problems with neural network based transformation models. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 2476-2481). IEEE.
5) Stojanoski, Z., Kalendar, M., & Gjoreski, H. Comparative Analysis of Machine Learning Models for Diabetes Prediction.
6) Albreiki, B., Zaki, N., & Alashwal, H. (2021). A systematic literature review of student'performance prediction using machine learning techniques. *Education Sciences*, *11*(9), 552.
7) Cunningham, P., & Delany, S. J. (2021). k-Nearest neighbour classifiers-A Tutorial. *ACM computing surveys (CSUR)*, *54*(6), 1-25.
8) Chapelle, O., Haffner, P., & Vapnik, V. N. (1999). Support vector machines for histogram-based image classification. *IEEE transactions on Neural Networks*, *10*(5), 1055-1064.
9) Hosmer, D. W., Taber, S., & Lemeshow, S. (1991). The importance of assessing the fit of logistic regression models: a case study. *American journal of public health*, *81*(12), 1630-1635.
10) Kotsiantis, S. B. (2013). Decision trees: a recent overview. *Artificial Intelligence Review*, *39*, 261-283.