ORDINAL REGRESSION WITH A TABULAR WINE QUALITY MODELS TEAM PROJECT

Name	ID
ZAYYAM FIDA	21032133

MY Role and Work

Data pre-processing specialist in the ordinal regression project utilizing a tabular wine quality dataset, my primary responsibility revolves around ensuring the data's quality and organization. This involves handling missing values and outliers, which entails identifying and addressing any data points that are incomplete or deviate significantly from the expected patterns. Additionally, I am responsible for scaling numerical features to ensure that they are on a comparable scale, enabling fair comparison and accurate analysis. I also handle categorical variables by appropriately encoding or transforming them into numerical representations, ensuring their compatibility with the regression model. Lastly, I divide the dataset into separate training and testing subsets, which enables us to evaluate the model's performance on unseen data. By executing these data pre-processing tasks, I contribute to creating a well-structured and standardized dataset that is conducive to accurate analysis and reliable results.

Missing Values

As a data pre-processing, my role is essential in ensuring the quality and reliability of the data used in a project. My responsibility for managing missing values and outliers involves identifying and handling any missing or extreme values that could affect the analysis. Moreover, Fig:1 Shows information on the dataset which is evidence of my work contributes to a project

```
] train_df.info()
  original_df.info()

  <class 'pandas.core.frame.DataFrame'>
  RangeIndex: 2056 entries, 0 to 2055
  Data columns (total 13 columns):
   #   Column                Non-Null Count  Dtype
  ---  ------                --------------  -----
   0   Id                    2056 non-null   int64
   1   fixed acidity         2056 non-null   float64
   2   volatile acidity      2056 non-null   float64
   3   citric acid           2056 non-null   float64
   4   residual sugar        2056 non-null   float64
   5   chlorides             2056 non-null   float64
   6   free sulfur dioxide   2056 non-null   float64
   7   total sulfur dioxide  2056 non-null   float64
   8   density               2056 non-null   float64
   9   pH                    2056 non-null   float64
   10  sulphates             2056 non-null   float64
   11  alcohol               2056 non-null   float64
   12  quality               2056 non-null   int64
  dtypes: float64(11), int64(2)
  memory usage: 208.9 KB
  <class 'pandas.core.frame.DataFrame'>
  RangeIndex: 1143 entries, 0 to 1142
```

Fig:1 Shows information of the dataset value

Separating The Dataset into Training and Testing Subsets Is Crucial For:

Evaluating model performance: It allows for an unbiased assessment of how well the model generalizes to unseen data, providing an accurate measure of its effectiveness.

Preventing overfitting: By testing the model on independent data, it helps detect overfitting, where the model memorizes the training data instead of learning patterns, leading to poor performance on new data.

Guiding model selection: The testing subset enables the comparison of different models, allowing the group to select the best-performing one for deployment, ensuring reliable and robust project outcomes.

Scaling numerical

Scaling numerical features is another key responsibility, where you transform variables to a common scale for fair comparisons and to avoid biases. Additionally, containing categorical variables involves converting them into numerical representations so they can be effectively used in models. Lastly, separating the dataset into training and testing subsets allows for accurate evaluation of model performance and generalizability. Fig2: Shows the training shape of a dataset

```
] full_train= pd.concat([train_df,original_df])
  full_train.drop('Id',axis=1, inplace= True)
  full_train.shape

  (3199, 12)
```

Fig 2: Shows the train shape of dataset

Min-Max scaling approach

the Min-Max scaling approach is important in data pre-processing for several reasons. It enables the normalization of variable scales, allowing for fair comparisons and avoiding the dominance of certain variables. It also preserves the relationships between data points, ensuring that patterns and structures in the data are maintained. Additionally, Min-Max scaling is compatible with certain algorithms, enhancing them. Fig 3: shows the minimax scalar approach of my contribution

```
[ ] #scale feature using minmax scaler approche
    X_train, X_test, y_train, y_test = train_test_split(scaled,y,train_size=0.8, test_size=0.2, random_state=1234)
    X_train.shape, X_test.shape

    ((2559, 11), (640, 11))
```

Fig 3: shows the minimax scalar approach

Learning Outcomes from Team

The significance of teamwork lies in its ability to foster collaboration, enhance problem-solving prowess, promote knowledge exchange, and facilitate the collective utilization of resources and abilities. By synchronizing varied perspectives and proficiencies through effective teamwork, individuals can propel themselves towards all-encompassing and triumphant outcomes. Moreover, teamwork nurtures a constructive and nurturing work milieu, where team members can acquire wisdom from one another and harmoniously contribute to a unified objective. The experience of collaborating within a team imparts a sense of contentment and fulfillment, as it cultivates personal growth and the attainment of shared ambitions.