Name: Pratik Manish Patil Div:D15C Roll:40

AIDS-I Assignment No: 2

Q.1: Use the following data set for question 1

82, 66, 70, 59, 90, 78, 76, 95, 99, 84, 88, 76, 82, 81, 91, 64, 79, 76, 85, 90

1. Find the Mean (10pts)
   To calculate the mean, first add up all the numbers in the dataset and then divide by the total number of values.

   Total Sum = 82 + 66 + 70 + 59 + 90 + 78 + 76 + 95 + 99 + 84 + 88 + 76 + 82 + 81 + 91 + 64 + 79 + 76 + 85 + 90 = 1691
   Count = 20

   Mean = Total Sum / Count = 1691 / 20 = **84.55**

   Therefore, the mean of the dataset is 84.55.

2. Find the Median (10pts)

   To find the median, I first need to arrange the data in ascending order:
   59, 64, 66, 70, 76, 76, 76, 78, 79, 81, 82, 82, 84, 85, 88, 90, 90, 91, 95, 99

   Since the dataset contains 20 values (an even number), the median is calculated by taking the average of the two middle numbers, which are the 10th and 11th values in the sorted list.
   Middle values: 81 and 82
   Median = (81 + 82) / 2 = 81.5
   Thus, the median of the dataset is 81.5.

3. Find the Mode (10pts)

   The mode is the value that appears most frequently in the dataset. Looking at the ordered list:

59, 64, 66, 70, 76, 76, 76, 78, 79, 81, 82, 82, 84, 85, 88, 90, 90, 91, 95, 99

The number 76 appears three times, which is more than any other number. Therefore, the mode of the dataset is 76.

4. Find the Interquartile Range (20pts)

The interquartile range (IQR) is the difference between the third quartile (Q3) and the first quartile (Q1).

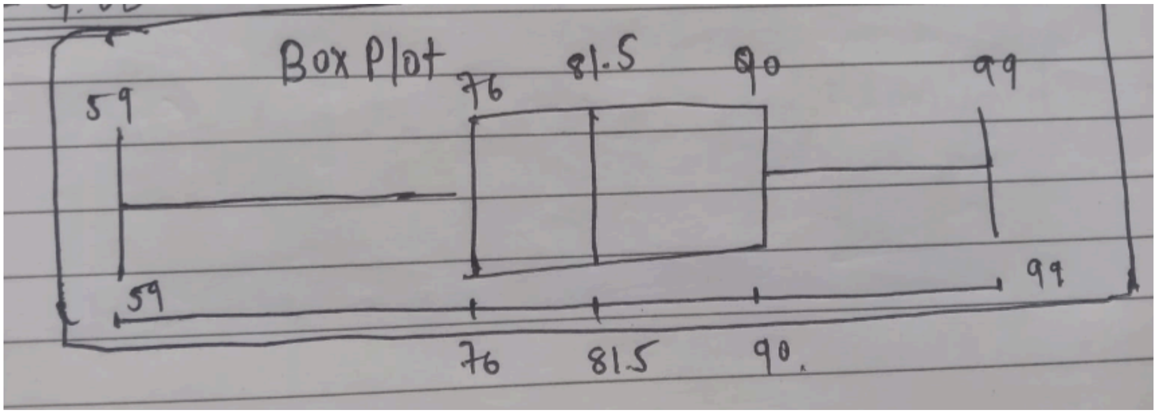
First, I need to find Q1 and Q3.

Ordered list: 59, 64, 66, 70, 76, 76, 76, 78, 79, 81, 82, 82, 84, 85, 88, 90, 90, 91, 95, 99

Q1, or the first quartile, represents the median of the lower half of the dataset. With 20 total values, the lower half includes the first 10 numbers. To find Q1, we calculate the average of the 5th and 6th values in the ordered list.

Lower half: 59, 64, 66, 70, **76, 76**, 76, 78, 79, 81
Q1 = (76 + 76) / 2 = 76

Q3, or the third quartile, is the median of the upper half of the dataset. Since the upper half consists of the last 10 values, **Q3** is calculated by taking the average of the 15th and 16th numbers in the ordered list.
Upper half: 82, 82, 84, 85, 88, **90, 90**, 91, 95, 99
Q3 = (90 + 90) / 2 = 90
IQR = Q3 - Q1 = 90 - 76 = 14
Thus, the interquartile range of the dataset is 14.



Q.2 1) Machine Learning for Kids 2) Teachable Machine

1. For each tool listed above:

    - Identify the target audience
    - Discuss the use of this tool by the target audience
    - Identify the tool's benefits and drawbacks
2. **1) Machine Learning for Kids**

    Target Audience:
    Machine Learning for Kids is primarily aimed at children and educators seeking an easy and engaging introduction to machine learning. It is designed for beginners with little to no background in coding or AI concepts.

    Tool Usage:
    The platform enables children to train machine learning models through intuitive, child-friendly interfaces. They can create models that recognize text, images, sounds, or numbers. These models can then be integrated into projects using Scratch or Python, allowing kids to build interactive games and applications that respond to their trained inputs. This hands-on approach helps learners grasp the fundamentals of machine learning by actively participating in the process.

Benefits:

- User-Friendly Design: Complex machine learning ideas are broken down into simple, understandable steps suitable for young learners.
- Interactive Learning: Integration with Scratch and Python adds a fun, creative layer to the educational experience.
- Educational Impact: The tool nurtures computational thinking and provides a foundational understanding of AI and machine learning.

Drawbacks:

- Limited Depth: Its simplified nature may not support more advanced machine learning projects or exploration.
- Platform Dependency: Requires integration with external platforms like Scratch or Python, which may need extra setup or guidance.

3. **2) Teachable Machine**

    - **Target Audience:**Teachable Machine is intended for a broad spectrum of users, including students, educators, artists, and hobbyists. It's designed for anyone interested in experimenting with machine learning in a simple, accessible way—no coding required.
    - **Use of the Tool:** Teachable Machine makes it easy to build machine learning models for image, sound, and pose recognition. Users can train models directly in their browser using live input from a webcam or microphone, or by uploading files. These models can be exported and embedded into websites, applications, or physical computing projects, making it a powerful tool for creative and educational prototyping.

- **Benefits:**
  - **Beginner-Friendly:** Requires no programming skills, making it highly approachable.
  - **Rapid Prototyping:** Allows users to quickly create and test ML models.
  - **Multiple Export Formats:** Models can be downloaded in different formats suitable for use in web apps, mobile apps, or TensorFlow environments.
- **Drawbacks:**
  - **Limited Customization:** Lacks advanced configuration options and tuning controls.
  - **Dependent on Browser:** Training and performance may be limited by browser and system resources.

4. From the two choices listed below, how would you describe each tool listed above? Why did you choose the answer?

   - Predictive analytic
   - Descriptive analytic
5. **1) Machine Learning for Kids:** Predictive analytic

   - **Why?**
     Machine Learning for Kids teaches users to build models that **predict outcomes** based on input data (e.g., classifying text, images, or numbers). When a child trains a model and uses it in a Scratch or Python project, the model is essentially **predicting a category or result** based on what it has learned.

     So, it’s **predictive analytics** because it:

     - Uses historical data (training examples)
     - Makes future predictions (e.g., “Is this a happy or sad sentence?”)
   -

6. **2) Teachable Machine:** Predictive analytic

    ○ **Why?**
    Teachable Machine also focuses on training models that **predict or classify** new inputs. For example, after training it to recognize images of different objects or sounds, the model will **predict what category** a new input belongs to.

    Thus, it fits **predictive analytics** because:

    - It uses input data to train a model
    - The model predicts future or unknown data outcomes (e.g., “This image is a cat”)

7. From the three choices listed below, how would you describe each tool listed above? Why did you choose the answer?

    ○ Supervised learning
    ○ Unsupervised learning
    ○ Reinforcement learning
    ○

8. **1) Machine Learning for Kids:** Supervised learning

    ○ **Why?** Machine Learning for Kids uses **labeled data** during training. For example, when children train a model to recognize happy or sad text, they **label** examples as “happy” or “sad.” The model learns from these labeled examples to make predictions on new, similar inputs.

    This is the essence of **supervised learning**:

    - The model is trained on input-output pairs (e.g., image → label)
    - It learns to map inputs to known categories or values

9. **2) Teachable Machine:** Supervised learning

   ○ **Why?** Teachable Machine also relies on **labeled training data**. For example, a user might provide multiple images and label them as “dog,” “cat,” or “person.” The model learns from these labeled examples to classify new inputs correctly.

   This is **supervised learning** because:

   - Users explicitly label the training examples
   - The model is trained to associate input data with those labels

Q.3 Data Visualization: Read the following two short articles:

Read the article Kakande, Arthur. February 12. "What's in a chart? A Step-by-Step guide to Identifying Misinformation in Data Visualization." Medium

Read the short web page Foley, Katherine Ellen. June 25, 2020. "How bad Covid-19 data visualizations mislead the public." Quartz

Research a current event which highlights the results of misinformation based on data visualization. Explain how the data visualization method failed in presenting accurate information. Use newspaper articles, magazines, online news websites or any other legitimate and valid source to cite this example. Cite the news source that you found.

Current Event Highlighting Misinformation Through Data Visualization: Misrepresented NOAA Temperature Graph

Event Overview

A graph sourced from the National Oceanic and Atmospheric Administration (NOAA) was selectively cropped and presented on social media to falsely claim that the Earth has been experiencing a cooling trend, contradicting the established scientific consensus on human-caused global warming.1 The graph focused solely on temperature data from 2015 to 2022, omitting long-term historical data that clearly shows a warming trend. This misrepresentation was fact-checked by the Associated Press (AP).

How the Data Visualization Method Failed

- **Cherry-Picking Data:** The visualization deliberately showcased only a limited timeframe (2015-2022), which was chosen to exploit natural climate variability (El Niño and La Niña events) and create a false impression of cooling. This selective presentation ignored the broader 140-year temperature record, a clear example of cherry-picking data to support a predetermined narrative.
- **Lack of Context:** The graph failed to provide essential context by omitting the long-term temperature data from NOAA. This omission prevented viewers from understanding the true extent of global warming and the significance of the short-term fluctuations shown in the graph. Legitimate data visualizations should always include sufficient context to ensure accurate interpretation.
- **Misleading Trendline:** A black line was added to the graph to emphasize a minor downward trend within the selected timeframe. This visual element drew attention to a statistically insignificant fluctuation, further reinforcing the false narrative of global cooling and obscuring the overall warming trend.
- **Exploitation of Authority:** The NOAA logo was prominently displayed on the graph, lending a false sense of authority and scientific validity to the misleading data. This tactic exploited the credibility of a reputable scientific institution to promote misinformation.

Impact of the Misinformation

- The misleading visualization contributed to the spread of climate change denial, undermining public understanding of the severity and urgency of global warming.
- It fostered skepticism towards climate science and reputable scientific institutions like NOAA, potentially eroding public trust in evidence-based information.
- The misrepresentation could diminish public support for policies and actions aimed at mitigating climate change.

Lessons for Ethical Data Visualization

- **Provide Complete Context:** Always include sufficient background information and long-term data to ensure accurate interpretation.

- **Avoid Cherry-Picking:** Present a comprehensive view of the data, avoiding selective presentation that supports a specific narrative.[2]
- **Ensure Data Accuracy:** Verify the accuracy and reliability of the data sources and methodologies used.
- **Transparency:** Clearly label and explain any data manipulations or selections.

News Source

- Associated Press (AP) Fact Check: "Temperature graph misrepresented to deny climate change," authored by Sophia Tulp, published on January 19, 2023.

Q. 4 Train Classification Model and visualize the prediction performance of trained model

Required information

Data File: Classification data.csv

Class Label: Last Column

Use any Machine Learning model (SVM, Naïve Base Classifier)

Requirements to satisfy

Programming Language: Python

Class imbalance should be resolved

Data Pre-processing must be used

Hyper parameter tuning must be used

Train, Validation and Test Split should be 70/20/10

Train and Test split must be randomly done

Classification Accuracy should be maximized

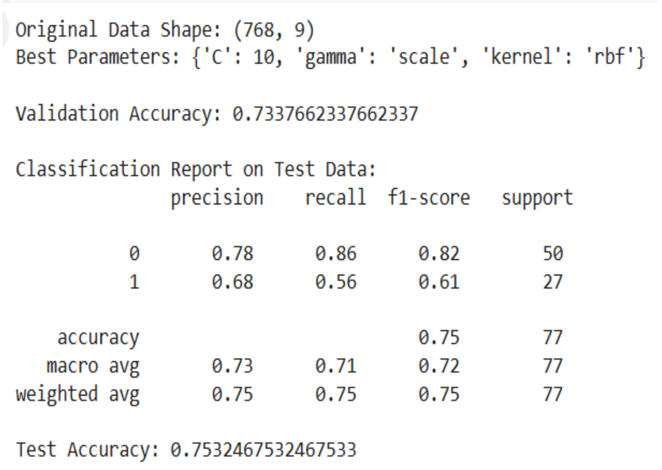
Use any Python library to present the accuracy measures of trained model

[Pima Indians Diabetes Database]

Explanation:

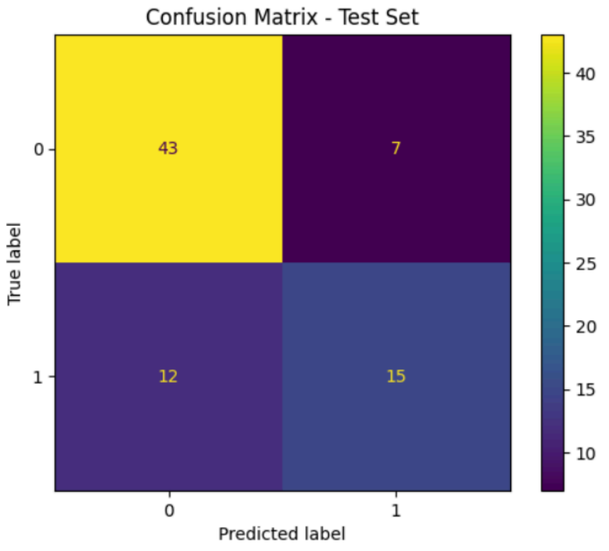
1. **Import Libraries:** data handling, modeling, and visualization.
2. **Load Data**: The Pima Indians Diabetes dataset is loaded using `pandas`.
3. **Data Preprocessing**: Zero values in health-related columns (e.g., Glucose, BMI) are replaced with NaN and filled with median values. Features are scaled using `StandardScaler`.
4. **Train-Validation-Test Split**: The dataset is split into 70% training, 20% validation, and 10% test sets with random shuffling and stratification.
5. **Class Imbalance Handling**: `SMOTE` is applied to the training set to balance the number of diabetic and non-diabetic cases.
6. **Hyperparameter Tuning**: `GridSearchCV` with 5-fold cross-validation is used to tune SVM hyperparameters (C, kernel, gamma) to maximize accuracy.
7. **Model Evaluation**: The best SVM model is evaluated on the test set using accuracy, classification report, and confusion matrix.
8. **Visualization**: The confusion matrix is visualized using `matplotlib` and `seaborn`.

With SVC :-

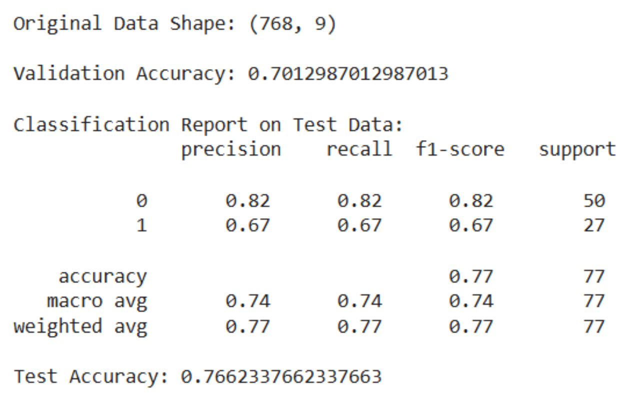


```
Original Data Shape: (768, 9)
Best Parameters: {'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}

Validation Accuracy: 0.7337662337662337

Classification Report on Test Data:
              precision    recall  f1-score   support

           0       0.78      0.86      0.82        50
           1       0.68      0.56      0.61        27

    accuracy                           0.75        77
   macro avg       0.73      0.71      0.72        77
weighted avg       0.75      0.75      0.75        77

Test Accuracy: 0.7532467532467533
```

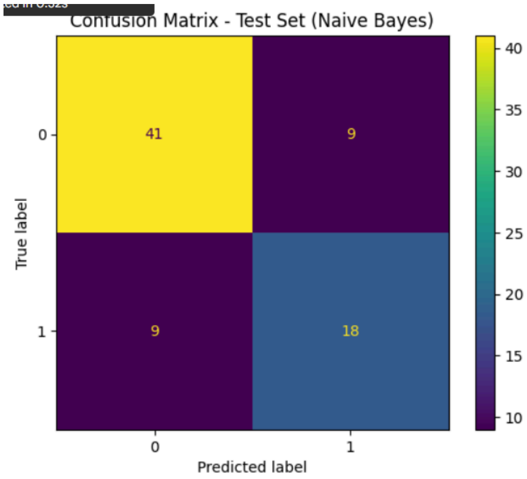
```
Test Accuracy: 0.7532467532467533
```



With Naive Bayes :-



```
Original Data Shape: (768, 9)

Validation Accuracy: 0.7012987012987013

Classification Report on Test Data:
              precision    recall  f1-score   support

           0       0.82      0.82      0.82        50
           1       0.67      0.67      0.67        27

    accuracy                           0.77        77
   macro avg       0.74      0.74      0.74        77
weighted avg       0.77      0.77      0.77        77

Test Accuracy: 0.7662337662337663
```



Q.5 Train Regression Model and visualize the prediction performance of trained model

Data File: Regression data.csv

Independent Variable: 1st Column

Dependent variables: Column 2 to 5

Use any Regression model to predict the values of all Dependent variables using values of Ist column.

Requirements to satisfy:

Programming Language: Python

OOP approach must be followed

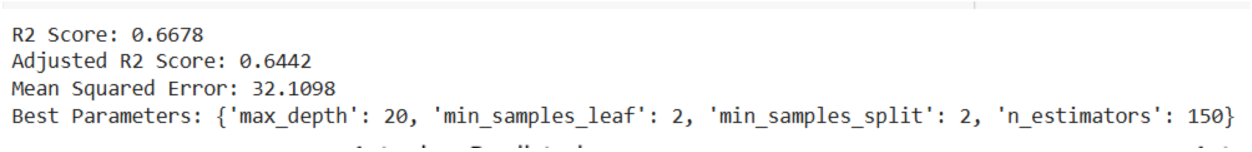
Hyper parameter tuning must be used

Train and Test Split should be 70/30

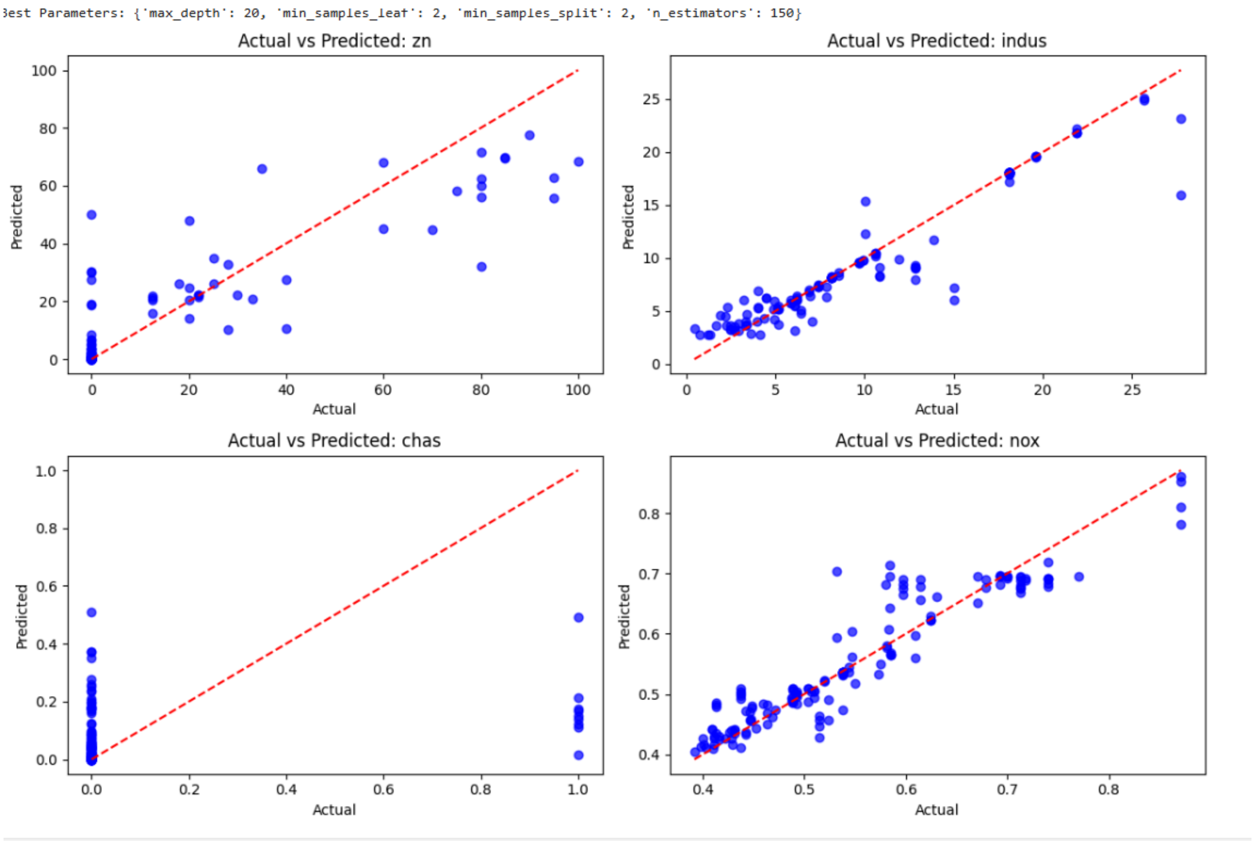
Train and Test split must be randomly done

Adjusted R2 score should more than 0.99

Use any Python library to present the accuracy measures of trained model



```
R2 Score: 0.6678
Adjusted R2 Score: 0.6442
Mean Squared Error: 32.1098
Best Parameters: {'max_depth': 20, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 150}
```



```
3est Parameters: {'max_depth': 20, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 150}
```

Q.6 What are the key features of the wine quality data set? Discuss the importance of each feature in predicting the quality of wine? How did you handle missing data in the wine quality data set during the feature engineering process? Discuss the advantages and disadvantages of different imputation techniques. (Refer dataset from Kaggle).

The wine quality data set from Kaggle primarily contains physicochemical measurements of wines and a quality score that reflects sensory evaluation. The key features (predictor variables) include:

- **Fixed Acidity:** Represents non-volatile acids that contribute to the wine's sour taste; it affects balance and structure.
- **Volatile Acidity:** Measures the presence of acetic acid; high levels can impart an unpleasant vinegar taste, negatively impacting quality.
- **Citric Acid:** Adds freshness and helps balance the wine's overall acidity; moderate amounts can enhance flavor.
- **Residual Sugar:** Indicates the unfermented sugar left in the wine; it plays a role in sweetness and overall flavor harmony, especially in white wines.
- **Chlorides:** Reflects the salt content; excessive levels can indicate poor sanitation or imbalance, thus reducing quality.
- **Free Sulfur Dioxide:** Acts as an antimicrobial and antioxidant; helps preserve wine flavor and stability, though excessive amounts may create off flavors.
- **Total Sulfur Dioxide:** Represents the overall amount used for preservation; important for shelf life but must be within safe sensory thresholds.
- **Density:** Closely related to the sugar content; contributes to the body and mouthfeel of the wine.
- **pH:** Provides an indication of wine acidity; optimal pH levels are necessary for microbial stability and overall flavor balance.
- **Sulphates:** Contribute to the wine's aroma and taste; they enhance complexity and act as an additional preservative measure.
- **Alcohol:** Affects body, viscosity, and flavor intensity; higher alcohol can balance high acidity in robust wines but may be overpowering if in excess.

Each of these features is vital because they collectively inform a model that predicts quality by capturing taste, balance, preservation, and overall sensory attributes. For example, while fixed and volatile acidity set the baseline for taste, residual sugar and citric acid can moderate harsh flavors; density and pH add to the body and stability, and alcohol contributes to both the structural and aromatic dimensions.

Missing Data Handling & Imputation Techniques:

In the wine quality data set, missing values may arise during data collection or preprocessing. Although many versions of this popular data set are complete, handling missing data is an essential step in feature engineering. The common imputation techniques include:

1. **Mean/Median Imputation:**
   • *Advantages:* Simple to implement and preserves the dataset size.
   • *Disadvantages:* Can distort distribution properties (especially with skewed data) and reduce variability.

2. **K-Nearest Neighbors (KNN) Imputation:**
   • *Advantages:* Uses local similarities to estimate missing values, which often leads to more accurate imputations.
   • *Disadvantages:* Computationally intensive and sensitive to the choice of distance metric and K value.

3. **Regression Imputation:**
   • *Advantages:* Leverages relationships among variables to predict missing values, potentially capturing complex dependencies.
   • *Disadvantages:* Can overfit and underestimate variability, as predicted values may cluster too tightly.

Each technique carries trade-offs; the best choice depends on the extent and pattern of missingness as well as the underlying data distribution. In practice, one might evaluate the "missing completely at random" (MCAR) assumption to decide if a simple mean/median imputation is sufficient, or if more sophisticated methods like KNN or regression imputation are required.

Overall, understanding the contribution of each feature helps build more accurate predictive models for wine quality, while careful handling of missing data ensures that the integrity of the model is maintained.