Assignment No. 2

Q.1 Data Set Analysis

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

1. Find the Mean (10 pts)

To find the mean:

- **Step 1:** Add all the numbers together.
  Sum = 82 + 66 + 70 + 59 + 90 + 78 + 76 + 95 + 99 + 84 + 88 + 76 + 82 + 81 + 91 + 64 + 79 + 76 + 85 + 90 = **1,621**
- **Step 2:** Divide the sum by the number of values (20).
  Mean = 1,621 ÷ 20 = 81.05

2. Find the Median (10 pts)

To find the median, first sort the data:
Sorted Data:
59, 64, 66, 70, 76, 76, 76, 78, 79, 81, 82, 82, 84, 85, 88, 90, 90, 91, 95, 99

- Since there are 20 numbers (even count), the median is the average of the 10th and 11th values.
  - 10th value = 81
  - 11th value = 82

- **Median = (81 + 82) ÷ 2 = 81.5**

3. Find the Mode (10 pts)

Mode is the number(s) that appear most frequently.

- **Observation:** 76 appears 3 times.
- All other numbers appear fewer times.
- **Mode = 76**

4. Find the Interquartile Range (20 pts)

Step 1: Order the data (already done):
59, 64, 66, 70, 76, 76, 76, 78, 79, 81, 82, 82, 84, 85, 88, 90, 90, 91, 95, 99

Step 2: Find Q1 (Lower Quartile)

- Lower half (first 10 numbers):
  59, 64, 66, 70, 76, 76, 76, 78, 79, 81
- Q1 is the average of the 5th and 6th numbers:
  Q1 = (76 + 76) ÷ 2 = 76

Step 3: Find Q3 (Upper Quartile)

- Upper half (last 10 numbers):
  82, 82, 84, 85, 88, 90, 90, 91, 95, 99
- Q3 is the average of the 5th and 6th numbers:
  Q3 = (88 + 90) ÷ 2 = 89

Interquartile Range (IQR) = Q3 - Q1 = 89 - 76 = 13

Q.2 Machine Learning Tools Analysis

1) Machine Learning for Kids

- **Target Audience:**
  Primarily designed for children and young learners (typically K–12), as well as educators introducing machine learning concepts in a classroom environment.
- **Use by Target Audience:**
  - Students use this tool to create simple machine learning projects such as games, chatbots, or recognition systems using visual programming (e.g., Scratch) combined with machine learning models.
  - Teachers use it to teach basic ML concepts in a hands-on and accessible way.
- **Benefits:**
  - User-friendly and educational interface
  - Encourages creative and experimental learning
  - Integrates well with Scratch and other block-based coding environments
  - Suitable for classroom settings
- **Drawbacks:**
  - Limited complexity — not suitable for advanced ML applications
  - Requires guidance from educators for effective learning
  - Less control over detailed algorithm parameters

2) Teachable Machine

- **Target Audience:**
  Geared toward a broader audience, including beginners, educators, hobbyists, and creators with little coding background.
- **Use by Target Audience:**
  - Users train models to recognize images, sounds, or poses using examples.
  - Often used in classrooms, prototypes, or art/interactive installations.

- **Benefits:**
    - Extremely simple and fast to use
    - No coding required
    - Can export models to TensorFlow or use them in apps and websites
    - Supports multiple input types (image, audio, pose)
- **Drawbacks:**
    - Limited scalability for large or complex datasets
    - Limited customization of model architecture
    - Performance can vary depending on training quality

2. Choice: Predictive Analytics

- **Machine Learning for Kids:**
  Predictive Analytic – Students use labeled data to train a model and then predict outcomes (e.g., classifying text or images based on training data).
- **Teachable Machine:**
  Predictive Analytic – This tool allows users to train a model using labeled examples and then make predictions based on new inputs (image, audio, or pose).

3. Choice: Supervised Learning

- **Machine Learning for Kids:**
  Supervised Learning – It uses labeled datasets to train models (e.g., categorizing objects or responses), which is a characteristic of supervised learning.
- **Teachable Machine:**
  Supervised Learning – The user provides examples with labels (e.g., "Class 1: Dog", "Class 2: Cat"), and the model learns to distinguish between them— a classic case of supervised learning.

Q.3 Data Visualization Analysis

1. **"What's in a chart? A Step-by-Step Guide to Identifying Misinformation in Data Visualization."**

- **Dual Nature of Data Visualization:**
  Kakande explains that the process of turning raw data into visual elements, such as charts, graphs, and infographics, can serve two distinct purposes. On one hand, well-crafted visualizations illuminate the data, helping the viewer quickly identify trends and patterns. On the other hand, poorly designed or intentionally misleading visuals can distort the information. This duality means that both creators and consumers of data visualizations must be vigilant about the underlying data and design choices.
- **Techniques for Misinformation:**
  The article outlines several common strategies by which data can be misrepresented:
    - **Truncated Graphs:** By deliberately not starting the y-axis at zero or by exaggerating the scale, a graph can make small differences look significant, or large differences appear minimal.

- **Misusing Color Scales:** Color is a powerful element in visual design. When color scales are inconsistently applied or swapped (for example, using red to indicate a positive trend when it is generally associated with a warning), it can lead to misinterpretation of the data.
- **Improper Pie Charts:** Pie charts are meant to represent parts of a whole. However, when the data segments do not sum up to 100% (or when the chart is otherwise poorly constructed), it can give a false sense of the data's structure or even its total quantity.

- **Responsibilities of Designers and Viewers:**
  Kakande stresses that the responsibility for accurate data visualization lies on two sides:
  - **Designers:** They must ensure that each chart or graph is purpose-driven—that is, it should be designed to display the most relevant data without exaggeration or omission. Designers should include clear annotations, reliable scales, and a well-chosen color palette that faithfully reflects the underlying data.
  - **Viewers:** Audiences should not accept visual information at face value. They are encouraged to critically assess key components such as axes, labels, and scales. Understanding the context behind the data and the visualization itself will help prevent misinterpretation or deception.

2. "How bad Covid-19 data visualizations mislead the public." – Quartz

- **Challenges in Early Pandemic Communications:**
  During the initial phases of the COVID-19 pandemic in the United States, state public health departments felt immense pressure to rapidly disseminate information. In their haste, many of these departments produced data visualizations that were intended to quickly inform the public about the spread and impact of the virus. However, the urgency sometimes led to oversimplified or cluttered graphics that inadvertently spread misinformation.
- **Specific Examples of Flawed Visualizations:**
  Foley points out several cases where poor visualization choices led to public confusion:
  - **Alabama's Visualizations:**
    The charts released by Alabama were criticized for presenting only snapshot data with cluttered numbers. Instead of providing an indication of trends over time (which is essential for understanding the progression of the outbreak), these visuals displayed data in a way that obscured meaningful analysis. Additionally, the use of pie charts contributed to confusion since pie charts can be difficult to interpret when too many small segments are involved.
  - **Arkansas' Approach:**
    In Arkansas, the visualizations often lacked sufficient context. For example, charts depicting the prevalence of preexisting health conditions among COVID-19 patients showed very low percentages on a scale of 0–100%. Without additional context, such as the actual number of cases or comparison with related conditions, the charts gave the misleading impression that the impact of those conditions was negligible. This minimized the public's understanding of the high risk certain patient populations faced.
  - **Arizona's Dashboard:**
    One of the charts featured in Arizona omitted a y-axis altogether. Additionally, the use of non-uniform color gradients contributed to a visual similarity between statewide data and data from a much smaller county. This poor design choice led viewers to incorrectly infer that the magnitude of COVID-19 cases was similar across

vastly different geographic areas, thereby distorting public perception of the actual risk levels.

Current Event Example: Misleading Voter Registration Visualization in Michigan for the 2024 Election

Overview:
In the lead-up to the 2024 U.S. election, social media posts began circulating an infographic claiming that Michigan had 500,000 more registered voters than eligible residents. Supporters of this narrative used the graphic to argue that there was widespread voter fraud in the state. The visualization was widely shared by influential figures as well as by prominent news aggregators, despite lacking proper context.

How the Data Visualization Method Failed:

- **Inappropriate Comparison Without Context:**
  The infographic showed two large bars side by side:
  - **Left Bar:** Representing the total number of registered voters (approximately 8.4 million).
  - **Right Bar:** Representing the number of Michigan residents of voting age (around 7.9 million).

    At first glance, the 500,000 difference appears alarming; however, the chart failed to explain that the total registered voter count includes both *active* voters and *inactive* voters (who have not voted in recent cycles but remain on the rolls by law). Many states maintain more registered voter records than there are active voters, due to legal retention practices and delayed removal processes.

- **Omission of Critical Data Segmentation:**
  A more accurate visualization would have broken down the "Registered Voters" category into two parts:
  - **Active Voters:** Individuals who have voted recently and are fully engaged.
  - **Inactive Voters:** Those who have not participated in the last few elections but remain registered because voter purges are subject to strict legal guidelines.

    By not segmenting the data, the infographic made it look as though the extra 500,000 registrations were fraudulent rather than a normal artifact of voter list maintenance.
- **Graphical Misrepresentation:**
  The design used an unqualified bar comparison (without annotations or supplemental notes) that encouraged a "quick glance" reading. Without labels or a clear key explaining that inactive voters legally remain on the list, audiences were likely to interpret the figure as evidence of intentional ballot stuffing or voter registration manipulation.

Impacts of the Misleading Visualization:
The misleading graphic helped fuel a narrative of voter fraud by implying an imbalance between registered voters and eligible voters. This misinterpretation undermined confidence in the electoral process and was amplified by prominent social media figures, influencing public debate in a politically polarized environment.

Graphical Representation (Description):
Imagine an infographic featuring:

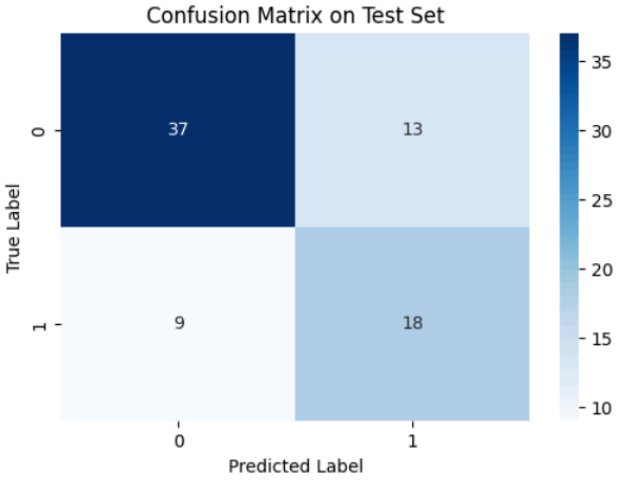
- **Two Side-by-Side Bars:**
  - **Left Bar (Total Registered Voters):** Labeled “8.4 Million Registered Voters” with an arrow pointing upward.
  - **Right Bar (Voting-Age Citizens):** Labeled “7.9 Million Eligible Voters.”

- **A Noticeable Gap:** A highlighted section indicating a difference of 500,000 voters.
- **Missing Annotation:** No distinction is made between active voters (who regularly vote) and inactive voters (who, while registered, do not vote).
  An effective redesign would add a third section to the “Registered Voters” bar that splits it into active versus inactive votes, along with explanatory footnotes detailing state laws regarding voter registration retention.

Q. 4 Train Classification Model

Output :

```
Validation Accuracy: 0.7662

Classification Report (Validation):
              precision    recall  f1-score   support

           0       0.86      0.77      0.81       100
           1       0.64      0.76      0.69        54

    accuracy                           0.77       154
   macro avg       0.75      0.76      0.75       154
weighted avg       0.78      0.77      0.77       154

Test Accuracy: 0.7143

Classification Report (Test):
              precision    recall  f1-score   support

           0       0.80      0.74      0.77        50
           1       0.58      0.67      0.62        27

    accuracy                           0.71        77
   macro avg       0.69      0.70      0.70        77
weighted avg       0.73      0.71      0.72        77
```



Q.5 Train Regression Model

Output :

```
WARNING: Not all dependent variables achieved an Adjusted R2 > 0.99.
Dependent variable 'relwt': Adjusted R2 = 0.0820
Dependent variable 'glufast': Adjusted R2 = 0.5946
Dependent variable 'glutest': Adjusted R2 = 0.6435
Dependent variable 'steady': Adjusted R2 = -0.1078
Dependent variable 'insulin': Adjusted R2 = 0.5300
Dependent variable 'group': Adjusted R2 = 0.8826
```

Q.6 Wine Quality Data Set Overview?

Key Features and Their Importance

1. **Fixed Acidity**
    - **Definition:** Non-volatile acids (e.g., tartaric, malic).
    - **Importance:** Contributes to sour taste and wine stability; indirectly impacts balance and quality.
2. **Volatile Acidity**
    - **Definition:** Mainly acetic acid, which evaporates readily.
    - **Importance:** High levels can cause off-odors; lower levels generally indicate higher quality.
3. **Citric Acid**
    - **Definition:** Naturally occurring acid in wine.
    - **Importance:** Enhances flavor complexity and adjusts overall acidity.
4. **Residual Sugar**
    - **Definition:** Sugar remaining after fermentation.
    - **Importance:** Influences sweetness, body, and balance of the wine.
5. **Chlorides**
    - **Definition:** Measures dissolved salts.
    - **Importance:** Excess may impart a saline taste; balanced levels indicate good winemaking.
6. **Free Sulfur Dioxide**
    - **Definition:** Active sulfur dioxide that prevents spoilage.
    - **Importance:** Preserves freshness but excess can harm taste.
7. **Total Sulfur Dioxide**
    - **Definition:** Sum of free and bound sulfur dioxide.
    - **Importance:** Reflects overall sulfur management and wine stability.
8. **Density**
    - **Definition:** Related to sugar and alcohol concentrations.
    - **Importance:** Acts as an indirect measure of residual sugar and alcohol.
9. **pH**
    - **Definition:** Measures wine acidity/basicity.
    - **Importance:** Critical for microbial stability and overall taste balance.
10. **Sulphates**
    - **Definition:** Compounds that aid antimicrobial and antioxidant properties.
    - **Importance:** Enhance aroma and preserve flavor integrity.
11. **Alcohol**
    - **Definition:** Produced during fermentation.
    - **Importance:** Determines body, strength, and is often directly linked to quality.
12. **Quality**
    - **Definition:** Sensory score (0–10) from expert tasters.
    - **Importance:** Serves as the output variable in predictive models.
13. **Id**
    - **Definition:** Unique sample identifier.
    - **Importance:** Useful for tracking but not predictive; typically dropped in modeling.

Handling Missing Data in Feature Engineering

Step 1: Identifying Missing Data

- Use methods such as df.isnull().sum() and visual tools to detect gaps.

Step 2: Imputation Techniques

- **Mean/Median Imputation:**

    - *Advantages:* Quick and simple.
    - *Disadvantages:* May distort distribution and underestimate variance.
- **KNN Imputation:**
    - *Advantages:* Considers feature relationships for improved estimates.
    - *Disadvantages:* More computationally intensive, sensitive to parameter choice.