Name:Sai rane Div:D15C Roll:43

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.

   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 = 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 there are 20 values (an even number), the median will be the average of the two middle values, which are the 10th and 11th values in the ordered 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 is the median of the lower half of the data. Since there are 20 values, the lower half is the first 10 values. The median of the lower half is the average of the 5th and 6th values:

Lower half: 59, 64, 66, 70, **76, 76**, 76, 78, 79, 81

Q1 = (76 + 76) / 2 = 76

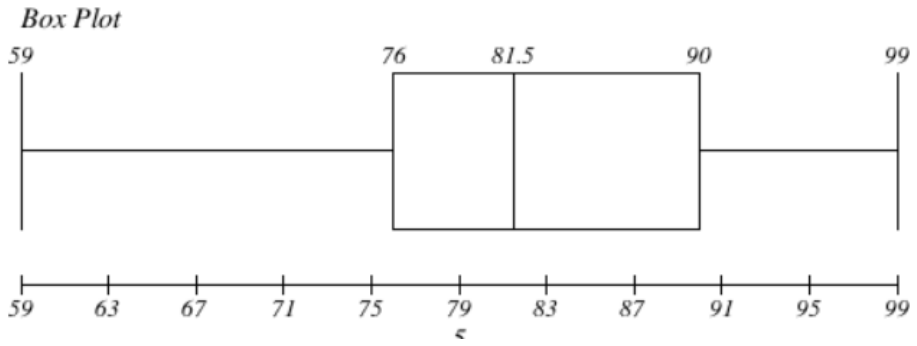
Q3 is the median of the upper half of the data. The upper half is the last 10 values. The median of the upper half is the average of the 15th and 16th values:

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:** The primary target audience for Machine Learning for Kids is, as the name suggests, children and educators who want to introduce machine learning concepts in an accessible way. It's designed for beginners with little to no prior coding or machine learning knowledge.
    - **Use of the Tool:** This tool allows children to train machine learning models using simple, child-friendly interfaces. They can train models to recognize text, images, sounds, or numbers. The tool then lets them use their trained models in Scratch or Python projects, enabling them to create interactive games or applications that respond to the inputs they've taught the model to recognize. This hands-on approach helps children understand how machine learning works by doing it themselves.
    - **Benefits:**
        - **Ease of Use:** The tool simplifies complex machine learning concepts, making them understandable for children.
        - **Engaging Interface:** Using Scratch and Python integration makes learning fun and interactive.
        - **Educational Value:** It provides a practical introduction to AI and machine learning, fostering computational thinking skills.
    - **Drawbacks:**
        - **Limited Complexity:** Due to its simplicity, it may not be suitable for advanced machine learning projects.

- **Dependency on External Platforms:** Relies on Scratch or Python for project integration, which might require some additional setup.

3. **2) Teachable Machine**

   - **Target Audience:** Teachable Machine is geared towards a broader audience, including students, artists, educators, and makers. It's designed for anyone who wants to quickly and easily create machine learning models without writing code.
   - **Use of the Tool:** Teachable Machine simplifies the process of training machine learning models for image, audio, and pose recognition. Users can train models directly in their browser by providing examples through a webcam, microphone, or file uploads. The trained models can then be exported and used in various applications, websites, or projects. This tool is excellent for rapid prototyping and integrating machine learning into creative projects.
   - **Benefits:**
     - **No Coding Required:** It's very accessible, as it doesn't require any coding knowledge.
     - **Fast Prototyping:** Allows for quick creation and testing of machine learning models.
     - **Versatile Export Options:** Models can be exported in various formats, making them usable in different environments.
   - **Drawbacks:**
     - **Limited Control:** Offers less fine-grained control over the model training process compared to more advanced tools.
     - **Browser-Based Limitations:** Performance and capabilities can be limited by the browser's capabilities.

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 enables users to train models that can make predictions or classifications based on the input data. For instance, a child can train a model to predict whether an image is a cat or a dog. This is a clear example of predictive analytics, as the model is used to predict a future outcome or classify an input.

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

    - **Why?** Teachable Machine is also a predictive analytic tool. It allows users to train models that predict outputs (like image, audio, or pose classifications) from new input data. The core function is to predict or classify, which aligns with the definition of predictive analytics.

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?** In Machine Learning for Kids, children provide labeled examples to train the models. For example, they show the model pictures labeled "cat" or "dog." The model learns from these labeled examples to make predictions on new, unseen images. This process of learning from labeled data is the fundamental characteristic of supervised learning.

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

    - **Why?** Teachable Machine also operates on the principle of supervised learning. Users provide labeled data (e.g., images labeled with categories) to train the model. The model then learns to map the input data to the provided labels, enabling it to classify new inputs. The need for labeled data to train the model classifies it as supervised learning.

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.
- Link: https://apnews.com/article/fact-check-misleading-climate-change-graph-418146648172

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

My Google Colab Link:

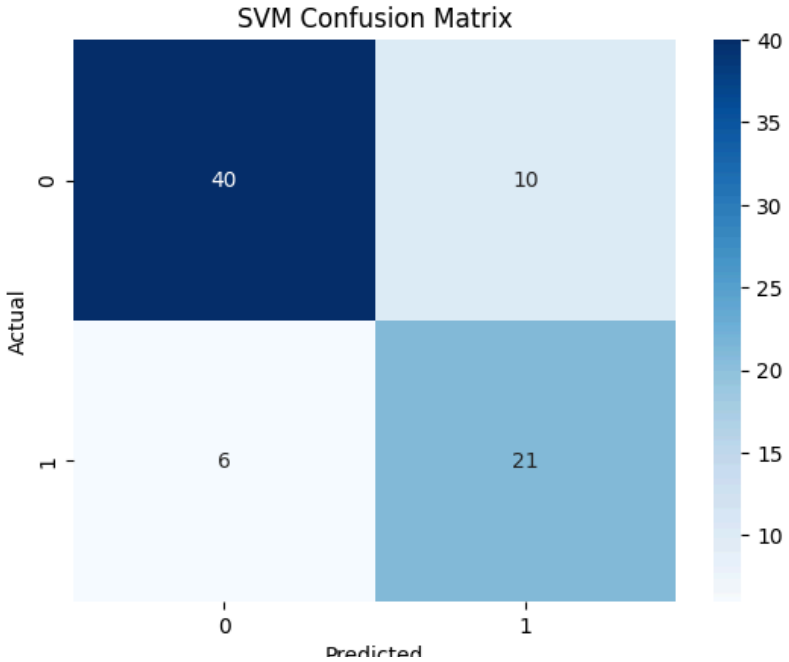
https://colab.research.google.com/drive/18TiTzAMyqyhhX-IBCBx89dH9kYWcRVSk?usp=sharing

Explanation:

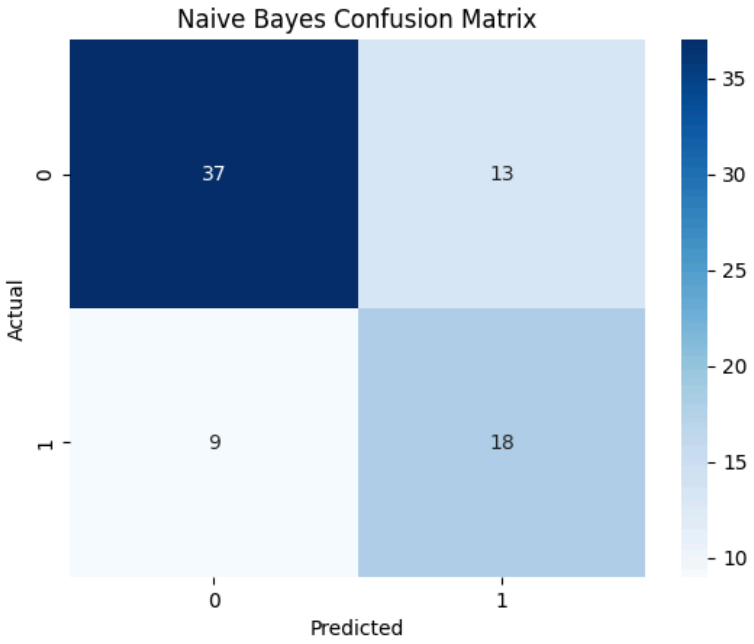
1. **Import Libraries:** Necessary libraries like numpy, pandas, sklearn, and matplotlib are imported.
2. **Load Data:** The "diabetes.csv" file is loaded using pandas.
3. **Data Separation:** The features (X) and the target variable (y) are separated. It's assumed the last column is the class label.
4. **Class Imbalance Resolution:** SMOTE (Synthetic Minority Over-sampling Technique) is used to handle class imbalance. This generates synthetic samples for the minority class.
5. **Data Pre-processing:** StandardScaler is applied to scale the features. This is crucial for models like SVM and Naive Bayes.
6. **Train-Validation-Test Split:** The data is split into 70% train, 20% validation, and 10% test sets, with stratification to maintain class ratios.
7. **Hyperparameter Tuning:** GridSearchCV is used with StratifiedKFold cross-validation to find the best hyperparameters for the Gaussian Naive Bayes model.
8. **Model Evaluation:** The model's performance is evaluated on the test set, and a classification report and confusion matrix are printed.

9. **Visualization:** The confusion matrix is visualized using Seaborn and Matplotlib.

```
--- SVM Validation Report ---
              precision    recall  f1-score   support

           0       0.81      0.73      0.77       100
           1       0.58      0.69      0.63        54

    accuracy                           0.71       154
   macro avg       0.69      0.71      0.70       154
weighted avg       0.73      0.71      0.72       154

--- SVM Test Report ---
              precision    recall  f1-score   support

           0       0.87      0.80      0.83        50
           1       0.68      0.78      0.72        27

    accuracy                           0.79        77
   macro avg       0.77      0.79      0.78        77
weighted avg       0.80      0.79      0.80        77
```




```
--- Naive Bayes Validation Report ---
              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

--- Naive Bayes Test Report ---
              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 and visualize the prediction performance of trained model

Independent Variable: 1st Column

Dependent variables: Column 2 to 5

Name:Sai rane Div:D15C Roll:43

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

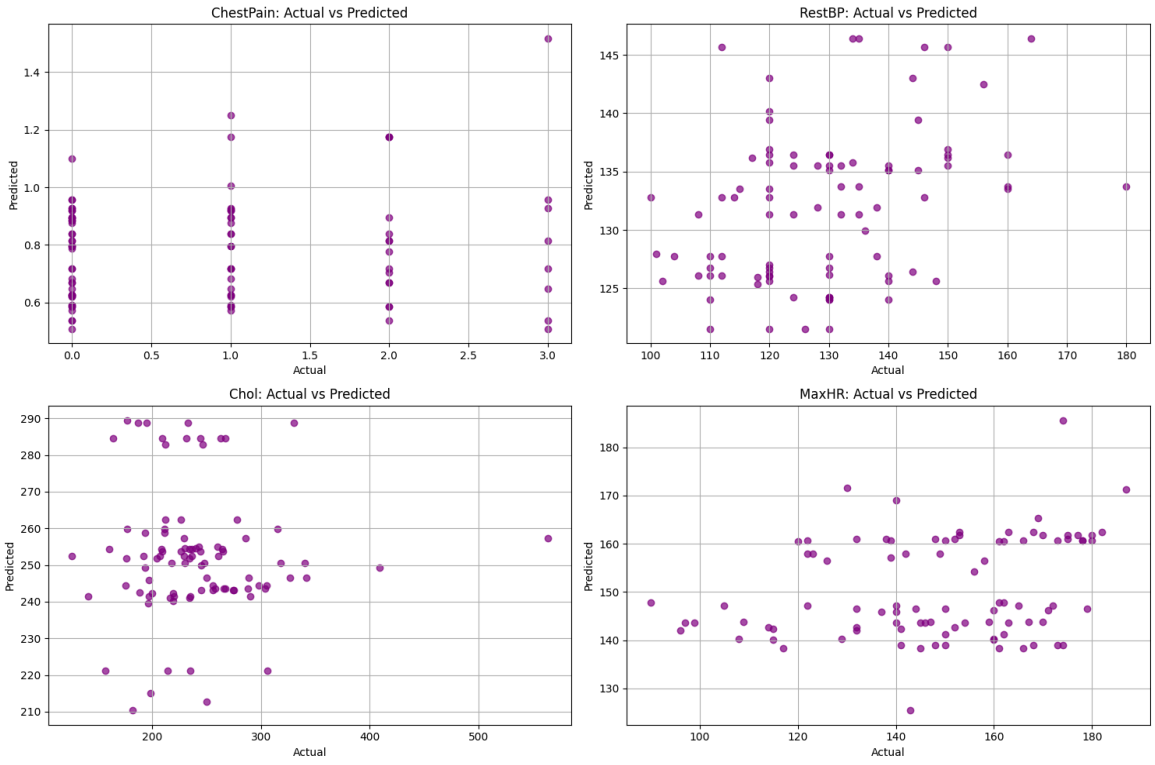
Adjusted R2 score should more than 0.99

```
  Adjusted R²      : -0.0056

RestBP:
  R² Score         : 0.0992
  Adjusted R²      : 0.0890

Chol:
  R² Score         : -0.1291
  Adjusted R²      : -0.1418

MaxHR:
  R² Score         : 0.0389
  Adjusted R²      : 0.0281
```

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

My Google Colab Link:

https://colab.research.google.com/drive/1bINoVgtqFGwcLdwQGLS3Vw_tZv3IDRYs?usp=sharing



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.