Name: Shivam.R.Prajapati **Div: D15C** **Roll No: 41**

Assignment No: 2

Q.1] Use the following 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
2. Find the Median
3. Find the Mode
4. Find the Interquartile range

Soln:

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

1. Mean :

$$\text{Mean} = \frac{\sum x_i}{n}$$

$$\text{Mean} = \frac{(82+66+70+59+90+78+76+95+99+84+88+76+82+81+91+64+79+76+85+90)}{20}$$

$$= \frac{1611}{20}$$

Mean = 80.55

2. Median:

Sort values: 59, 64, 66, 70, 76, 76, 76, 78, 79, 81, 82, 82, 84, 85, 88, 90, 90, 91, 95, 99
Total values 20 that is Total Even Values are there

$$\text{Median} = \frac{x_{(n/2)} + x_{(n/2)+1}}{2}$$

Even number of values → average of 10th and 11th:

$$\text{Median} = \frac{81+82}{2} = 81.5$$

Median = 81.5

3. Mode:

The mode is the value that appears most frequently in a dataset.
Most frequent value = 76
Mode = 76

4. Interquartile Range (IQR)

$$\mathrm{IQR} = Q3 - Q1$$

Q1 = 25th percentile = average of 5th and 6th

Q1 = $\frac{76+76}{2} = 76$

Q3 = 75th percentile = average of 15th and 16th

Q3 = $\frac{88+90}{2} = 89$

IQR = Q3−Q1

= 89−76

IQR =13

Therefore,
Mean = 80.55 ,
Median = 81.5 ,
Mode = 76 ,
IQR = 13

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

- **Predictive analytic**
- **Descriptive analytic**

3. 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**

Soln:

The following Machine Learning Tools Comparison:

1. Machine Learning for Kids
2. Teachable Machine

Name: Shivam.R.Prajapati **Div: D15C** **Roll No: 41**

1] Machine Learning for Kids:

- Target Audience are School students (ages 8 to 16), beginners, and educators teaching AI/ML in schools or basic courses.
- Users can train ML models using Text, Images, Numbers
- It connects with platforms like Scratch and Python, allowing users to build interactive projects like:
    - A chatbot that detects positive/negative messages
    - An app that recognizes fruits from images
- Example :A student can upload labeled photos of cats and dogs and then use Scratch to make a game that guesses whether a new image is a cat or dog.
- Advantages:
    - User-friendly interface, great for young learners.
    - Integrates ML with visual programming (Scratch).
    - Encourages creative projects and experimentation.
    - No coding required (but optional Python use is available).
    - Cloud-based, accessible from browsers.
- Limitations:
    - Limited to basic models; lacks depth for real-world ML applications.
    - Accuracy is low compared to professional tools.
    - Minimal control over algorithm types, hyperparameters, or data preprocessing.
    - Not suitable for large datasets or complex models.

2] Teachable Machine:

- Target Audiences are General public, students, educators, hobbyists, and even artists.
- It is use for:
    - Creating models by training with webcam/audio/images.
    - Exporting the model to use in websites or apps.
    - Because No coding required.
- Advantages:
    - Extremely simple to use with a few clicks.
    - Fast real-time training with immediate feedback.
    - Supports exporting trained models to real applications.
    - Great for interactive demos, art installations, and education.
- Limitations:
    - No deep customization or control over the model architecture.
    - No preprocessing options (e.g., normalization).
    - Limited dataset size and simple structure = low accuracy on complex problems.
    - Cannot handle text or numerical data.

Name: Shivam.R.Prajapati **Div: D15C** **Roll No: 41**

2. Choosing Predictive or Descriptive Analytic:

Tool	Type	Reason
Machine Learning for Kids	Predictive Analytic	It uses trained models to predict categories or outputs from inputs.
Teachable Machine	Predictive Analytic	It predicts labels for new input data (like image or sound classification).

We have not chosen descriptive because Descriptive analytics explains what happened in the past using statistics and visualization. These tools instead predict outcomes using new input data.

3. Choosing Type of Learning.

Tool	Learning Type	Reason
Machine Learning for Kids	Supervised Learning	It uses labeled data (e.g., text labeled as positive/negative) to train.
Teachable Machine	Supervised Learning	Users provide labeled examples for training (e.g., face = "happy").

We have not chosen Unsupervised or Reinforcement because

1) These tools don't discover hidden patterns or reward strategies on their own.
2) They rely on explicit labels provided by the user, which defines 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.

Name: Shivam.R.Prajapati **Div: D15C** **Roll No: 41**

Soln:

Case Study: Misleading COVID-19 Vaccine Death Visualizations

In early 2024, misleading COVID-19 vaccine-related visuals began circulating widely on social media platforms. These visuals suggested that vaccinated individuals in England experienced more deaths than those who were unvaccinated between July 2021 and May 2023. At a glance, the charts looked alarming and prompted concern—but they lacked critical context, leading to serious misinterpretation. (Source: Reuters, March 21, 2024 – "Misleading data used to claim COVID vaccines do more harm than good")

Where the Visualization Went Wrong:

1. **Lack of Proportional Context**:
   The visuals displayed the **total number of deaths** in each group (vaccinated vs. unvaccinated) without acknowledging that the **vast majority of England's population had been vaccinated** during this period. Naturally, with more people in the vaccinated group, the raw number of deaths would be higher—but that doesn't indicate higher risk. Without adjusting for group size, the comparison becomes misleading.

2. **Missing Essential Background**:
   The charts **did not explain how effective vaccines are**, nor did they highlight the **difference in group sizes**. This omission led many viewers to assume that vaccines were ineffective or dangerous, which is a misinterpretation based purely on incomplete data presentation.

3. **Ignoring Death Rates**
   Instead of showing **deaths per 100,000 people** (which allows for fair comparison between differently sized groups), the charts showed raw death counts. This skewed the perceived risk, concealing the fact that **unvaccinated individuals were statistically at greater risk of dying** from COVID-19 during this time frame.

Clarifying the Misrepresentation:

Experts reviewed the same data with proper adjustments—especially accounting for **population size**—and revealed a different story. According to the UK's Office for National Statistics (ONS), **death rates among vaccinated individuals were significantly lower** than among unvaccinated individuals. The issue wasn't with the data itself, but rather how it was **visualized and interpreted**. When data lacks context, it can easily fuel confusion or spread misinformation.

Name: Shivam.R.Prajapati **Div: D15C** **Roll No: 41**

Conclusion:

This case highlights how **critical it is to design data visualizations responsibly**. Charts and graphs that omit proportionality, context, or clear labeling can mislead audiences—even if the data is technically correct. In public health, such misinterpretations can have dangerous consequences by eroding public trust. As noted by experts like Arthur Kakande and Katherine Ellen Foley, both creators and consumers of data must be **vigilant, critical, and context-aware** to ensure that visuals truly inform rather than mislead.

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

Soln:

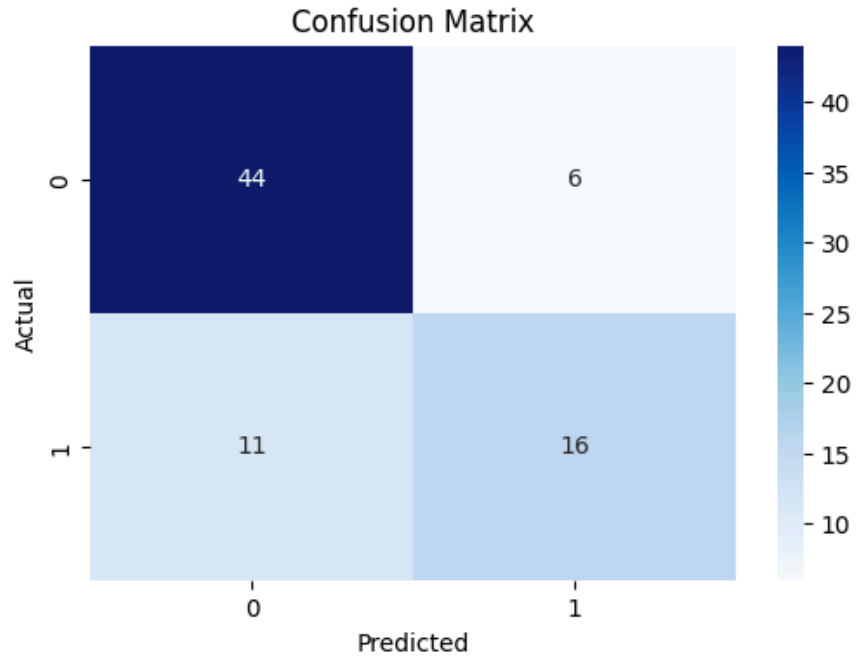
- **Pregnancies**: Number of times the patient has been pregnant.
- **Glucose**: Plasma glucose concentration after a 2-hour oral glucose tolerance test.
- **BloodPressure**: Diastolic blood pressure (mm Hg).
- **SkinThickness**: Triceps skin fold thickness (mm).
- **Insulin**: 2-hour serum insulin (mu U/ml).
- **BMI**: Body Mass Index (weight in kg / height in m^2).
- **DiabetesPedigreeFunction**: A function that scores the likelihood of diabetes based on family history.
- **Age**: Patient age in years.

Name: Shivam.R.Prajapati **Div: D15C** **Roll No: 41**

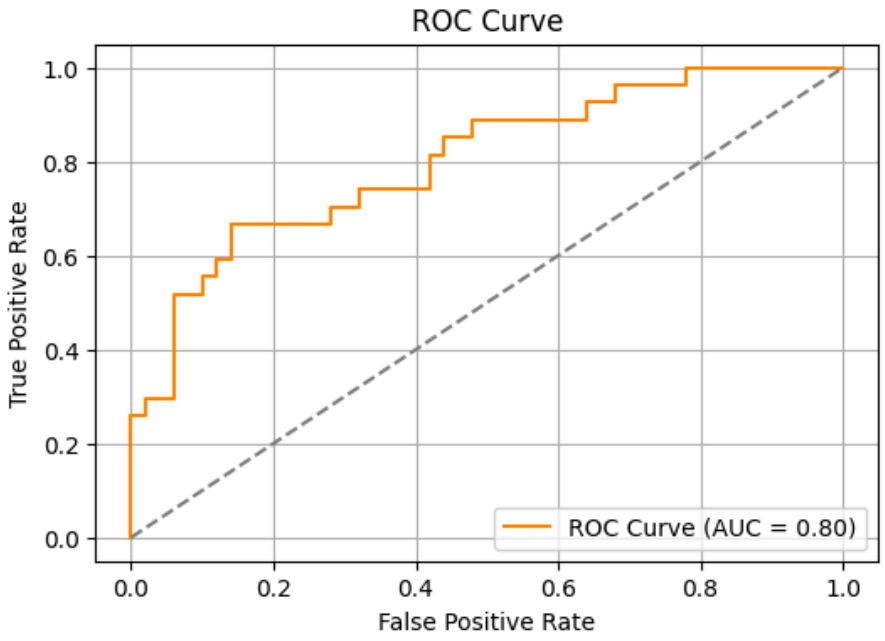
Model: **Support Vector Machine (SVM)**

Result:

```
Data loaded successfully.
Preprocessing complete.
Model training complete with best parameters.
Best Parameters: {'svm__C': 10, 'svm__gamma': 0.1, 'svm__kernel': 'rbf'}

Accuracy on Test Set: 0.7792

Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.88      0.84        50
           1       0.73      0.59      0.65        27

    accuracy                           0.78        77
   macro avg       0.76      0.74      0.75        77
weighted avg       0.77      0.78      0.77        77
```



The SVM model achieved an accuracy of **77.92%** on the test set, with better performance in detecting non-diabetic cases (class 0) than diabetic ones (class 1). Although precision and recall for class 1 are lower, the overall classification is reasonably balanced. With hyperparameter tuning, the model shows good potential for early diabetes prediction.



The ROC analysis yielded an **AUC of 0.80**, indicating that the model demonstrates good discriminative ability between diabetic and non-diabetic cases. This suggests that, on average, the classifier correctly distinguishes between the two classes 80% of the time. While the performance is acceptable for practical applications, further refinement in feature engineering or model tuning could be explored to enhance predictive accuracy.

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 the 1st 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**

https://github.com/Sutanoy/Public-Regression-Datasets

https://raw.githubusercontent.com/selva86/datasets/master/BostonHousing.csv

Name: Shivam.R.Prajapati **Div: D15C** **Roll No: 41**

URL:
https://archive.ics.uci.edu/ml/machine-learning-databases/00477/Real%20estate%20valuation%20data%20set.xlsx

(Refer any one)

Soln:

Dataset: Dry_Bean_Dataset

Dataset features:

- **Area**: Total pixel count inside the bean region.
- **Perimeter**: Distance around the bean boundary.
- **MajorAxisLength**: Length of the longest axis of the bean.
- **MinorAxisLength**: Length of the shortest axis of the bean.
- **AspectRatio**: Ratio of major to minor axis.
- **Eccentricity**: How elongated the bean is.
- **ConvexArea**: Number of pixels in the convex hull of the bean.
- **EquivDiameter**: Diameter of a circle with the same area as the bean.
- **Extent**: Ratio of bean area to bounding box area.
- **Solidity**: Ratio of bean area to convex hull area.
- **roundness**: Circularity of the bean shape.

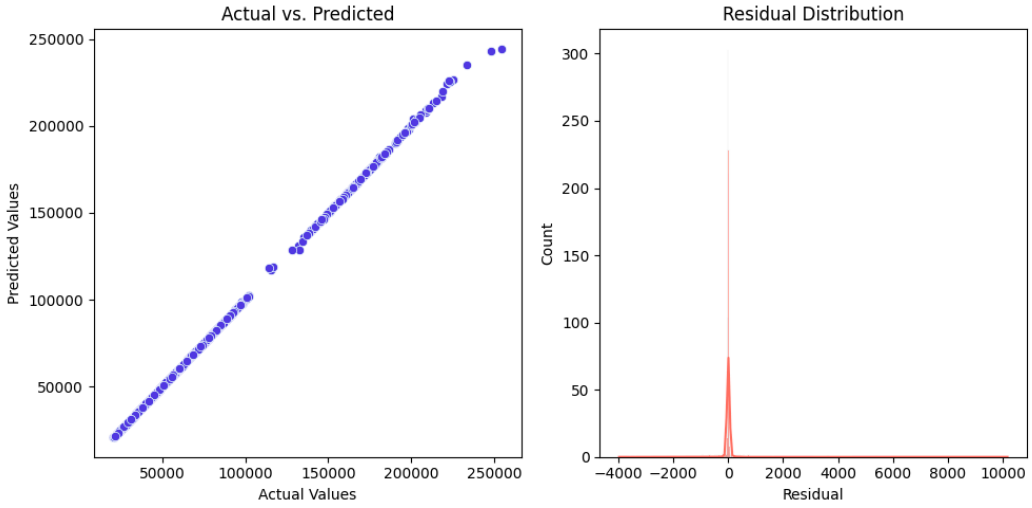
Compactness, ShapeFactor1, ShapeFactor2, ShapeFactor3, ShapeFactor4: Geometrical shape descriptors.

Model: **RandomForestRegressor**

Here we are predicting Area based on other features

Result:

```
Best parameters:  {'max_depth': 10, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 50}
R2 Score: 0.9999
Adjusted R2 Score: 0.9999
✅ Model meets the required Adjusted R² score.
```



The regression model, tuned with optimal hyperparameters, achieved an **R^2 and Adjusted R^2 score of 0.9999**, indicating an **excellent fit** with the data. It successfully satisfies the requirement of an Adjusted R^2 score above **0.99**, demonstrating high predictive accuracy and generalization capability for the dry bean dataset.

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).

Soln:

Main Features and Their Importance:

1. **Fixed Acidity:** Represents non-volatile acids (tartaric, malic, citric) that remain relatively stable. Higher fixed acidity can contribute to tartness and affect overall balance.
2. **Volatile Acidity:** Primarily acetic acid content, which at high levels can give an unpleasant vinegar taste. Low volatile acidity is generally preferred for higher quality wines.
3. **Citric Acid:** Can add freshness and flavor to wines. It contributes to the wine's citrus character and can enhance perceived freshness.
4. **Residual Sugar:** Affects the sweetness of the wine. Different wine styles have different optimal levels, making this feature important but contextual.
5. **Chlorides:** Salt content in wine, which can influence taste perception. Excessive chlorides can negatively impact quality.
6. **Free Sulfur Dioxide:** Acts as a preservative and antioxidant. Appropriate levels help preserve wine quality while excessive amounts can create unpleasant aromas.
7. **Total Sulfur Dioxide:** Sum of free and bound forms of SO_2. Important for preservation but can be detrimental to flavor when too high.

8. **Density:** Related to alcohol and sugar content. Provides information about the wine's body and can indicate fermentation completeness.
9. **pH:** Affects chemical stability and microbial control. Wines with balanced pH tend to be more stable and often higher quality.
10. **Sulphates:** Additives that contribute to SO_2 levels and act as preservatives. Moderate levels help wine stability.
11. **Alcohol:** Higher alcohol content is often associated with higher quality ratings, as it contributes to body and can enhance flavor perception.

Handling Missing Data in the Wine Quality Dataset:

Common Imputation Techniques and Their Trade-offs:

1. **Mean/Median/Mode Imputation**
   - **Advantages**: Simple, fast implementation; preserves the mean (when using mean imputation)
   - **Disadvantages**: Reduces variance; ignores relationships between features; can distort distributions
2. **K-Nearest Neighbors (KNN) Imputation**
   - **Advantages**: Accounts for relationships between features; better preserves data structure
   - **Disadvantages**: Computationally expensive for large datasets; sensitive to outliers; requires parameter tuning
3. **Regression Imputation**
   - **Advantages**: Maintains relationships between variables; can be more accurate than simpler methods
   - **Disadvantages**: May overfit; assumes linear relationships; can reduce variance
4. **Multiple Imputation**
   - **Advantages**: Accounts for uncertainty in missing values; maintains variability in the dataset
   - **Disadvantages**: Computationally intensive; more complex to implement
5. **Machine Learning Based Imputation** (Random Forest, etc.)
   - **Advantages**: Can capture complex non-linear relationships; often provides more accurate imputations
   - **Disadvantages**: Computationally expensive; risk of overfitting; requires careful validation