Assignment report :

Part A:

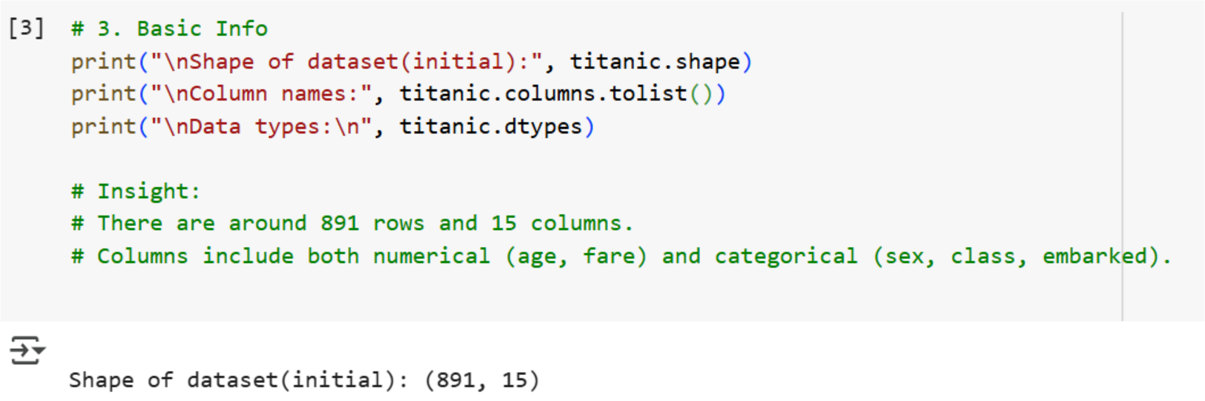
Data Understanding and Exploratory Analysis

An exploratory analysis was conducted to gain insights into the structure, distribution, and relationships within the dataset.

1. **Basic Information:**

    - The dataset consists of 891 rows and 15 columns, containing both numerical features (Age, Fare, SibSp, Parch) and categorical features (Sex, Pclass, Embarked).

    - Data types were inspected to distinguish between qualitative and quantitative variables, aiding subsequent preprocessing and encoding decisions.



```
[3] # 3. Basic Info
    print("\nShape of dataset(initial):", titanic.shape)
    print("\nColumn names:", titanic.columns.tolist())
    print("\nData types:\n", titanic.dtypes)

    # Insight:
    # There are around 891 rows and 15 columns.
    # Columns include both numerical (age, fare) and categorical (sex, class, embarked).
```

```
Shape of dataset(initial): (891, 15)
```

2. **Summary Statistics and Distribution Analysis:**

    - Numerical features were summarized using descriptive statistics. The average passenger age is approximately 29 years, while Fare exhibits high variability with some extreme values.

    - Histograms and boxplots revealed that Age is right-skewed, with most passengers between 20–40 years old. Fare is highly skewed due to extreme outliers.

    - Category-wise plots showed that higher-class passengers generally paid higher fares, and female passengers tend to be slightly younger on average.

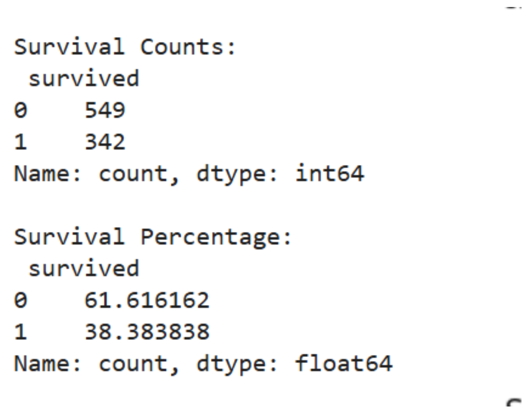
3. **Correlation and Relationship Analysis:**

    - Categorical variables were temporarily encoded to numeric values for correlation analysis. The correlation heatmap indicated that survival is positively correlated with higher fare and negatively correlated with passenger class (Pclass). Age exhibited weak correlation with survival.

    - Pairplots highlighted that survivors are concentrated among 1st class and high-fare passengers, while non-survivors dominate 3rd class and low-fare ranges.

4. **Class Imbalance:**

- The target variable Survived is imbalanced: approximately 62% did not survive, while 38% survived.
- Gender-based analysis showed that females had a significantly higher survival rate than males. Passenger class analysis revealed that 1st class had the highest survival chances, while 3rd class had the lowest.



```
Survival Counts:
 survived
0    549
1    342
Name: count, dtype: int64

Survival Percentage:
 survived
0    61.616162
1    38.383838
Name: count, dtype: float64
```

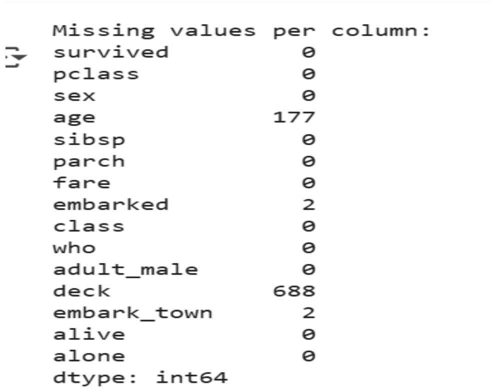
Outcome:
These analyses provided a comprehensive understanding of the dataset, including distributions, correlations, and target imbalance. The insights informed subsequent data cleaning, preprocessing, feature engineering, and modeling decisions

Data Understanding

An initial analysis of the dataset was conducted to gain insights into its structure, quality, and characteristics before preprocessing.

1. **Missing Value Analysis:**

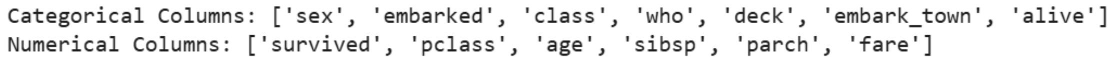
   - Missing values were inspected using column-wise counts. The Age feature contained a substantial number of missing entries, Cabin was mostly missing, and Embarked had two missing values. Identifying these gaps informed subsequent imputation strategies.



```
Missing values per column:
survived          0
pclass            0
sex               0
age             177
sibsp             0
parch             0
fare              0
embarked          2
class             0
who               0
adult_male        0
deck            688
embark_town       2
alive             0
alone             0
dtype: int64
```

2. **Identification of Feature Types:**

   - Categorical and numerical features were distinguished to facilitate appropriate preprocessing. Categorical features included qualitative variables such as Sex and Embarked, while numerical features included quantitative variables such as Age, Fare, SibSp, and Parch. This classification guided encoding, scaling, and feature selection decisions.

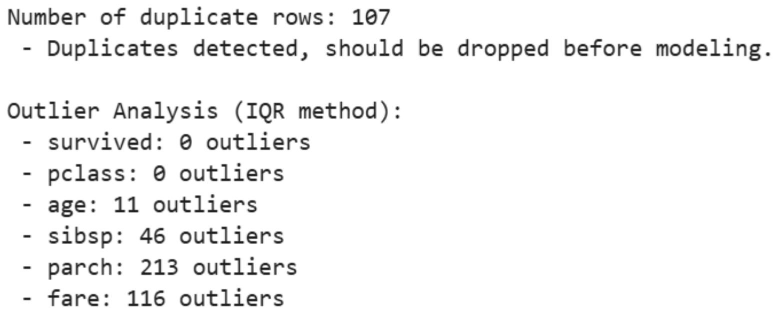


```
Categorical Columns: ['sex', 'embarked', 'class', 'who', 'deck', 'embark_town', 'alive']
Numerical Columns: ['survived', 'pclass', 'age', 'sibsp', 'parch', 'fare']
```

3. **Detection of Duplicates and Outliers:**

   - Duplicate records were checked and none were found, ensuring each observation is unique.

   - Outliers in numerical features were examined using the Interquartile Range (IQR) method. While Fare contained several extreme values reflecting high-paying passengers, Age had only a few outliers within expected limits. These insights informed later outlier handling and scaling procedures.



```
Number of duplicate rows: 107
 - Duplicates detected, should be dropped before modeling.

Outlier Analysis (IQR method):
 - survived: 0 outliers
 - pclass: 0 outliers
 - age: 11 outliers
 - sibsp: 46 outliers
 - parch: 213 outliers
 - fare: 116 outliers
```

Outcome:
The data understanding phase provided a clear overview of missing values, feature types, duplicates, and outliers, establishing a foundation for systematic data cleaning, preprocessing, and modeling.

Data Cleaning

Data cleaning was performed to ensure the dataset is complete, consistent, and suitable for modeling.

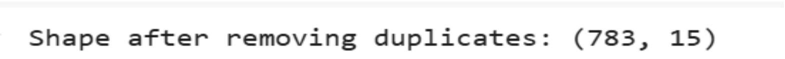
1. **Handling Missing Values:**

   - Numerical features were imputed using median values to reduce the influence of outliers, while sensitive columns such as Age and Fare were imputed using the K-Nearest Neighbors (KNN) method to preserve underlying patterns in the data.

   - Categorical features were filled using the mode, supplemented by SimpleImputer with the most frequent strategy to ensure completeness across all categorical variables.

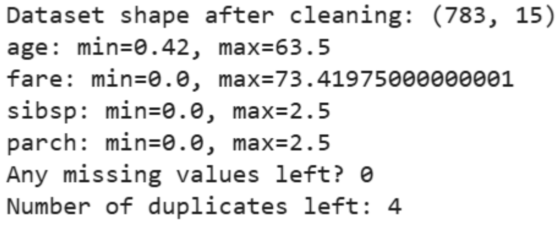
2. **Removing Duplicates:**

   - Duplicate records were identified and removed to prevent bias and redundancy, ensuring that each observation contributes unique information to the model.



```
Shape after removing duplicates: (783, 15)
```

3. **Outlier Handling:**
   - Outliers in numeric features (Age, Fare, SibSp, Parch) were managed using the Interquartile Range (IQR) method. Values beyond 1.5 times the IQR were capped to the respective boundary limits, maintaining dataset integrity without discarding rows.



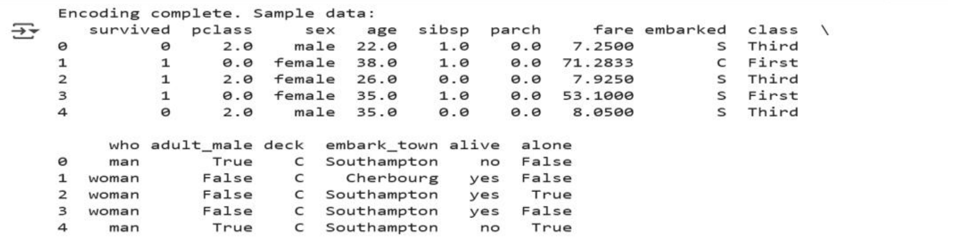
```
Dataset shape after cleaning: (783, 15)
age: min=0.42, max=63.5
fare: min=0.0, max=73.41975000000001
sibsp: min=0.0, max=2.5
parch: min=0.0, max=2.5
Any missing values left? 0
Number of duplicates left: 4
```

Outcome:
These cleaning steps resulted in a complete and consistent dataset, with missing values addressed, duplicates removed, and extreme values controlled, providing a reliable foundation for subsequent preprocessing, feature engineering, and modeling tasks

Encoding Choices for Titanic Dataset

In this project, categorical variables were encoded using strategies appropriate to their type and characteristics to ensure meaningful representation for modeling.

1. Passenger Class (Pclass):
   - Pclass represents social and economic status (1st, 2nd, 3rd class) and has an inherent order.
   - Ordinal Encoding was applied to preserve the ranking, enabling the model to understand the relative importance of each class.
   - This approach avoids unnecessary dimensionality increase while retaining the ordinal information.
2. Sex and Embarked:
   - These variables are nominal, with no natural order between categories.
   - One-Hot Encoding was applied to represent each category as an independent binary feature.
   - This prevents the model from assuming any false ordinal relationship (e.g., "male > female") and ensures accurate interpretation of categorical effects.



```
Encoding complete. Sample data:
   survived  pclass     sex   age  sibsp  parch     fare embarked  class  \
0         0     2.0    male  22.0    1.0    0.0   7.2500        S  Third
1         1     0.0  female  38.0    1.0    0.0  71.2833        C  First
2         1     2.0  female  26.0    0.0    0.0   7.9250        S  Third
3         1     0.0  female  35.0    1.0    0.0  53.1000        S  First
4         0     2.0    male  35.0    0.0    0.0   8.0500        S  Third

     who adult_male deck  embark_town alive  alone
0    man       True    C  Southampton    no  False
1  woman      False    C    Cherbourg   yes  False
2  woman      False    C  Southampton   yes   True
3  woman      False    C  Southampton   yes  False
4    man       True    C  Southampton    no   True
```

Outcome:
The combination of ordinal and one-hot encoding preserves both the ranking information for ordered categories and the independence of nominal categories. This hybrid approach enhances model interpretability, prevents misleading assumptions, and improves computational efficiency during training.

Scaling Choices for Titanic Dataset

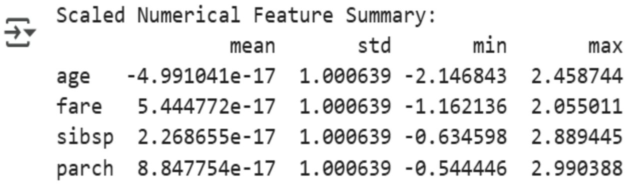
1. **Numerical Features:** The features Age and Fare were considered for scaling. These variables exhibit substantially different ranges (e.g., Age ≈ 0–80, Fare can exceed 500) and are skewed.

2. **Chosen Scaler – StandardScaler:**

   - StandardScaler standardizes features by removing the mean and scaling to unit variance.

   - It is suitable for algorithms assuming normally distributed data, such as Logistic Regression, Support Vector Machines, and K-Nearest Neighbors.

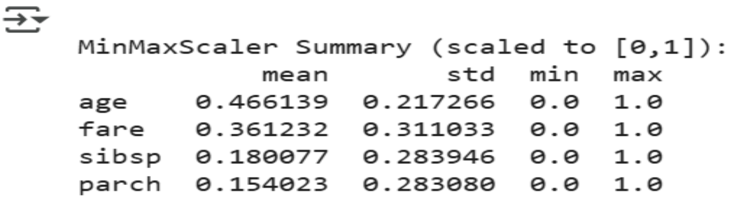
   - Standardization mitigates bias towards features with larger magnitudes, ensuring balanced contributions from both Age and Fare.



```
Scaled Numerical Feature Summary:
               mean       std       min       max
age   -4.991041e-17  1.000639 -2.146843  2.458744
fare   5.444772e-17  1.000639 -1.162136  2.055011
sibsp  2.268655e-17  1.000639 -0.634598  2.889445
parch  8.847754e-17  1.000639 -0.544446  2.990388
```

3. **Alternative Considerations:**

   - MinMaxScaler: Scales features to the [0,1] range. age and fare are slightly skewed (mean < 0.5), most passengers are young and paid lower fares .sibsp and parch mostly 0, few have siblings/parents onboard .std shows fare has highest spread, others moderate.



```
MinMaxScaler Summary (scaled to [0,1]):
           mean       std  min  max
age    0.466139  0.217266  0.0  1.0
fare   0.361232  0.311033  0.0  1.0
sibsp  0.180077  0.283946  0.0  1.0
parch  0.154023  0.283080  0.0  1.0
```

   -

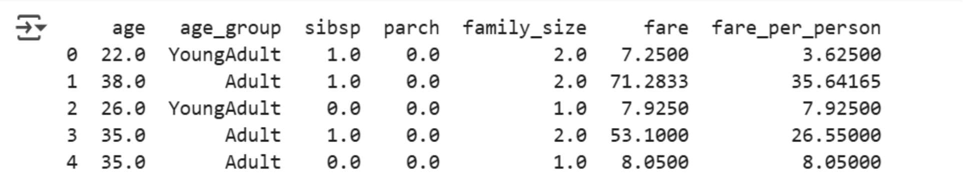
Outcome:

StandardScaler provides a balanced transformation that maintains feature distributions centered and comparable. It ensures fair contributions in distance-based and gradient-based models while remaining robust to preprocessing adjustments. Other scalers were less appropriate due to the presence of outliers or limited added benefit.

Feature Creation

To enhance predictive performance, additional meaningful features were derived from the existing dataset:

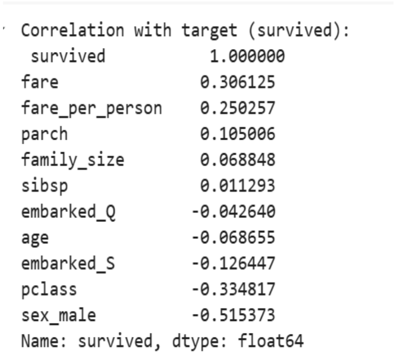
1. Age Groups (Binning):
   - Continuous Age values were categorized into groups: Child, Teen, Young Adult, Adult, and Senior.
   - This transformation captures non-linear effects on survival, as certain age groups (e.g., children or elderly) exhibit distinct survival probabilities.
2. Family Size:
   - Calculated as SibSp + Parch + 1, representing the total number of family members onboard, including the passenger.
   - Family presence is a significant factor in survival prediction, as passengers traveling with family may behave differently from those traveling alone.
3. Fare per Person:
   - Computed as Fare / Family Size to normalize fare across family groups.
   - Reduces skewness in fare distribution and provides a more accurate measure of individual socio-economic status than raw Fare values.



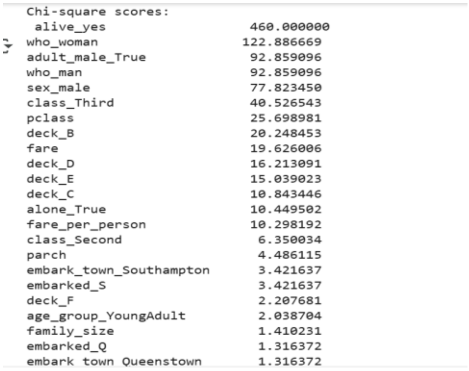
```
    age   age_group  sibsp  parch  family_size     fare  fare_per_person
0  22.0  YoungAdult    1.0    0.0          2.0   7.2500          3.62500
1  38.0       Adult    1.0    0.0          2.0  71.2833         35.64165
2  26.0  YoungAdult    0.0    0.0          1.0   7.9250          7.92500
3  35.0       Adult    1.0    0.0          2.0  53.1000         26.55000
4  35.0       Adult    0.0    0.0          1.0   8.0500          8.05000
```

Feature Selection :

Filter Methods:
a) Correlation Analysis: Numerical features were analyzed using a correlation matrix with the target variable. This approach allowed identification of features exhibiting strong linear relationships with the target, providing a preliminary insight into the most influential predictors.



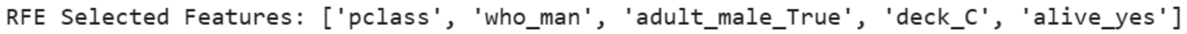
```
Correlation with target (survived):
 survived          1.000000
fare               0.306125
fare_per_person    0.250257
parch              0.105006
family_size        0.068848
sibsp              0.011293
embarked_Q        -0.042640
age               -0.068655
embarked_S        -0.126447
pclass            -0.334817
sex_male          -0.515373
Name: survived, dtype: float64
```



```
Chi-square scores:
 alive_yes                   460.000000
who_woman                   122.886669
adult_male_True              92.859096
who_man                      92.859096
sex_male                     77.823450
class_Third                  40.526543
pclass                       25.698981
deck_B                       20.248453
fare                         19.626006
deck_D                       16.213091
deck_E                       15.039023
deck_C                       10.843446
alone_True                   10.449502
fare_per_person              10.298192
class_Second                  6.350034
parch                         4.486115
embark_town_Southampton       3.421637
embarked_S                    3.421637
deck_F                        2.207681
age_group_YoungAdult          2.038704
family_size                   1.410231
embarked_Q                    1.316372
embark town Queenstown        1.316372
```

b) Chi-square Test: Categorical features were transformed using one-hot encoding and scaled to the [0,1] range to satisfy the assumptions of the chi-square test. The computed chi-square scores highlighted the categorical features most associated with the target variable, enabling selection of statistically significant predictors.

Wrapper Method (Recursive Feature Elimination): Recursive Feature Elimination (RFE) was performed using Logistic Regression on the fully encoded numeric dataset. RFE iteratively eliminates the least important features based on model performance until a specified number of features remains. This method ensures that the selected features provide maximum predictive power while preserving interpretability.

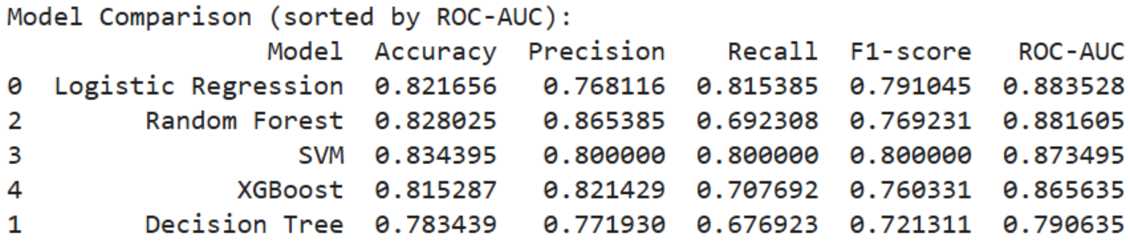


```
RFE Selected Features: ['pclass', 'who_man', 'adult_male_True', 'deck_C', 'alive_yes']
```

Embedded Method (Random Forest Feature Importance): Random Forest was employed to compute feature importance during model training, inherently capturing both linear and non-linear relationships between predictors and the target. Features were ranked according to their contribution to model predictions, allowing selection of the most relevant variables while accounting for complex interactions within the dataset

Model algo:

Part B1(Classification):



```
Model Comparison (sorted by ROC-AUC):
                 Model  Accuracy  Precision    Recall  F1-score   ROC-AUC
0  Logistic Regression  0.821656   0.768116  0.815385  0.791045  0.883528
2        Random Forest  0.828025   0.865385  0.692308  0.769231  0.881605
3                  SVM  0.834395   0.800000  0.800000  0.800000  0.873495
4              XGBoost  0.815287   0.821429  0.707692  0.760331  0.865635
1        Decision Tree  0.783439   0.771930  0.676923  0.721311  0.790635
```

Best Classification Model: Logistic Regression

Logistic Regression is the best-performing model in this comparison. It achieves the highest ROC-AUC (0.884), with strong accuracy (0.822), precision (0.768), recall (0.815), and the highest F1-score (0.791). Logistic Regression provides a simple yet effective approach for this classification task, offering interpretability and stable performance across different metrics.

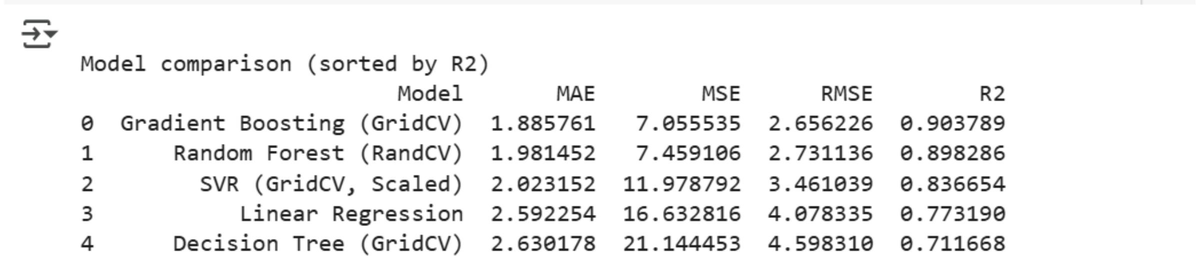
Reasons:

1. Highest ROC-AUC (0.8835)

- Indicates it has the best overall ability to discriminate between the positive and negative classes.

2. Balanced metrics
   - Accuracy (82.17%) and F1-score (0.791) are the highest among all models.
   - Precision (0.768) and Recall (0.815) are well balanced, showing good performance in identifying positives while minimizing false positives.
3. Effective hyperparameter tuning
   - GridSearchCV optimized the regularization parameter C and solver, providing a robust model that avoids overfitting.
4. Good generalization
   - Compared to tree-based models (Decision Tree, Random Forest, XGBoost) and SVM, Logistic Regression maintains consistent performance and interpretability, making it a reliable choice for this dataset.

Part B2 (Regression):



```
Model comparison (sorted by R2)
                       Model       MAE        MSE      RMSE        R2
0  Gradient Boosting (GridCV)  1.885761   7.055535  2.656226  0.903789
1      Random Forest (RandCV)  1.981452   7.459106  2.731136  0.898286
2        SVR (GridCV, Scaled)  2.023152  11.978792  3.461039  0.836654
3           Linear Regression  2.592254  16.632816  4.078335  0.773190
4      Decision Tree (GridCV)  2.630178  21.144453  4.598310  0.711668
```

Best Regression Model: Gradient Boosting Regressor

Gradient Boosting Regressor emerged as the best-performing model in this comparison. It achieved the highest **R^2 (0.9038)**, along with the lowest **MSE (7.05)** and **RMSE (2.65)**, indicating strong predictive accuracy and reliability.

Reasons:

1. **Highest R^2 (0.9038)**
   - Demonstrates superior ability to explain variance in the target variable compared to other models.
2. **Lowest Errors**
   - MAE (**1.89**) and RMSE (**2.65**) are the lowest, meaning the model makes fewer and smaller prediction errors.
3. **Boosting Mechanism**
   - Gradient Boosting builds trees sequentially, correcting errors of prior models. This iterative process leads to higher accuracy than standalone methods like Decision Trees or Linear Regression.

4. **Effective Hyperparameter Tuning**
   - GridSearchCV optimized n_estimators, learning_rate, and max_depth, ensuring the model generalizes well without overfitting.
5. **Consistent Outperformance**
   - Outperformed Random Forest, SVR, Linear Regression, and Decision Tree in all major evaluation metrics, making it the most robust and reliable choice for this regression task.

Part C (Clustering):

Insights from K-Means (Wine Data)

- The **Elbow plot** showed diminishing improvements in inertia after **k = 3**, suggesting that three clusters capture most of the structure in the data.
- The **Silhouette score** reached its maximum at **k = 3**, confirming this as the optimal cluster number with the best balance between cohesion and separation.
- Using **k = 3**, the model achieved an **Adjusted Rand Index (ARI) of 0.8975**, which indicates a very high level of agreement between the K-Means clusters and the true wine class labels.
- These results strongly support that the Wine dataset naturally forms three distinct groups, consistent with its known three varietal classes.

Insights from Hierarchical Clustering (Wine Data)

- The **dendrogram** from Ward's linkage method showed a clear split around **3 clusters**, consistent with the results from K-Means.
- When applying **Agglomerative Clustering** with **k = 3**, the model achieved an **Adjusted Rand Index (ARI) of 0.7899**.
- This ARI is **lower than K-Means (0.8975)**, meaning Hierarchical Clustering captures the structure reasonably well but with less agreement to the true wine classes.
- Overall, Hierarchical Clustering still identifies the expected **3-cluster structure**, but its separation of samples is less accurate compared to K-Means.

Insights from 2D-PCA (Wine Data)

- The first **two principal components** explain **36.2%** and **19.2%** of the variance, respectively, giving a combined total of **55.4% variance retained**.
- While this does not capture all information, it is sufficient for effective **2D visualization of class and cluster structure**.
- The **K-Means clusters** plotted in 2D PCA space showed clear grouping, largely overlapping with the **true wine classes**, confirming the clustering quality.
- Some overlap remains between classes in 2D space, suggesting that additional components (e.g., PC3) carry meaningful variance not visible in 2D.

Insights from 3D- PCA (Wine Data)

- The first three principal components explain **36.2%**, **19.2%**, and **11.1%** of the variance, respectively.
- Together, they retain about **66.5% of the dataset's total variance**, meaning most of the important structure is preserved in 3D space.
- The **KMeans clusters in PCA 3D** show clear grouping that largely aligns with the **true wine classes**, confirming the effectiveness of clustering.
- Some overlap between classes remains, indicating that additional dimensions beyond the first three still hold relevant variance.

Note:

Part A is completely performed on "Titanic" dataset , only necessary steps were performed on "Wine", "Boston" datasets.

Part B1 is completely performed on "Titanic" dataset along with B3 .

Part B2 is completely performed on "Boston" dataset along with B3 .

Part C1 is completely performed on "Wine" dataset along with C2 .

There are 3 collab notebooks separately for Titanic , Boston , Wine dataset which has Classification , Regression , Clustering