Problem 1

Task 1

Initial Visualisation of dataset was done.

Classification of Variables :

- Ordinal Variables : Pclass

- Nominal or Categorical Variables : Sex, Embarked, Survived

- Continuous Variables : Age, Family Size, Fare

Preprocessing :

- Handling of missing values : Missing values in age column were filled by mean of values corresponding to same class and sex. Cabin column was removed as it had 77% missing data and was not an important feature. Rows with missing Embarked information were removed because only 0.22% of the data was missing Embarked information.

- Merging of Sibsp and Parch : Both the columns were merged into a single column named Family Size.

- Checking for Outliers : Outliers were checked for Age, Fare and Family Size but were not considered because they represent elderly people, wealthy people and bigger families and are no threat to data.

- Categorical Encoding : Age and Embarked were encoded and Pclass was already encoded.

- Visualisation after preprocessing : Features were visualised using heat map and different plots.

Splitting of data : Data was split randomly using 70-20-10 split in train, validation and test data.

Task 2-8

Helper Functions :

- `calculateEntropy():` Function returns the entropy given the set of values.

- `entropySplit():` Function decides the best threshold value for splitting using information gain as the criteria.

- `bestAttribute():` Function gives the best attribute or feature on which split has to be applied using information gain and also return the corresponding threshold value and left-right datasets after splitting.

- `classify():` Classify the node as leaf node and make necessary changes.

- `class TreeNode():` Class contains node objects of the decision tree with necessary properties.

Implementation of Tree :

- `createTree():` Creates a tree with base conditions on minimum samples, maximum depth and minimum information gain.

- `infer():` Predict the survival and gives the corresponding accuracy.

Results :

Accuracy on Test Data : 86.67%
Accuracy on Validation Data : 83.05%
Accuracy on Train Data : 84.41%
Overall Accuracy : 83.13%
Precision: 0.89
Recall: 0.80
F1-Score: 0.84

Problem 2

Task 1-2

Data Exploration : Plotted a scatter plot between Sales and TV. Visualisation of statistical measures was done.

Preprocessing : No missing values were found in the data. Z-score normalisation was applied to the TV column.

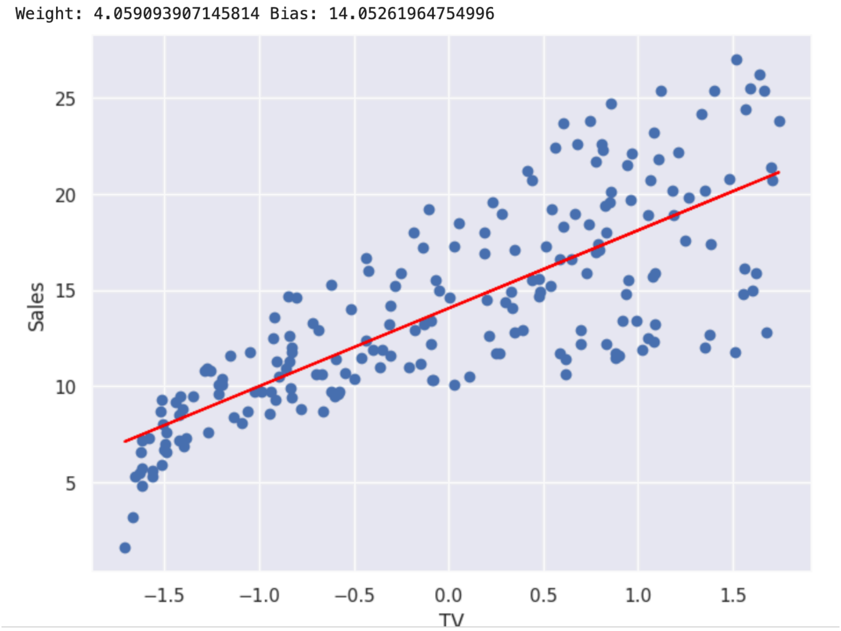
Splitting of data : 80-20 split was used to split the data into train and test datasets.

Task 3-4

Linear Regression Implementation :

`costFunction(), updateWeights()` and `linearRegression()` functions were implemented to calculate mean square error as cost and gradient descent to update weights.

Result :



```
Weight: 4.059093907145814 Bias: 14.05261964754996
```

Mean Absolute Error : 2.8365

Mean Square Error : 11.9920

Problem 3

Task 1-2

Data Exploration : Statistical measures were shown, heat map was used to show the correlation between features. Distribution of target variable was shown using density plot.

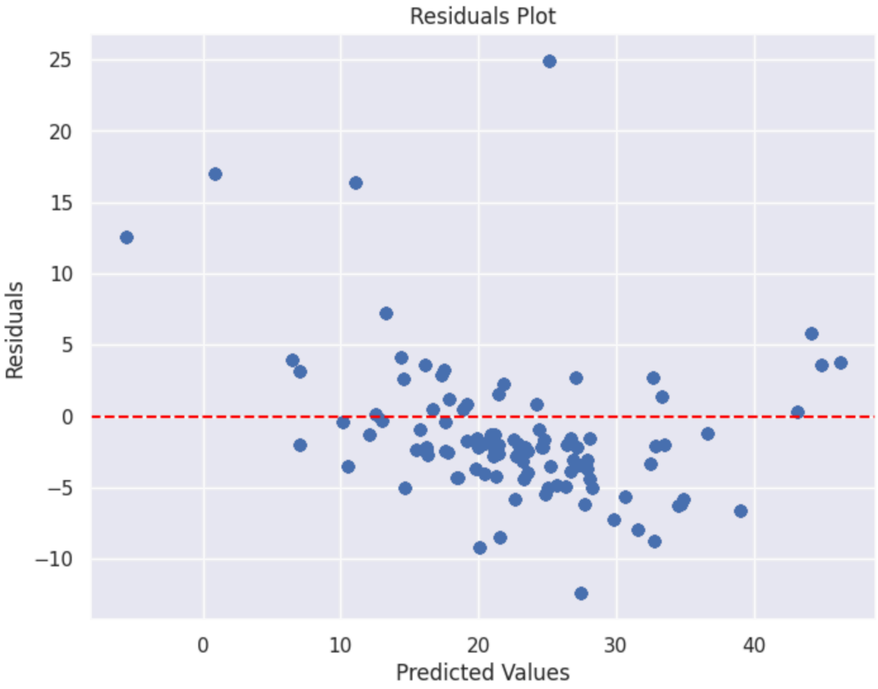
Preprocessing : Missing values were about 4% in some columns and were filled with the mean values of the corresponding columns.

Splitting of data : 80-20 split was used to split the data into train and test datasets.

Multiple Linear Regression Implementation :

`compute_cost(), featureScaling()` and `gradient_descent()` were used to calculate mean square error as cost, normalise the data and to update the weights respectively.

Result : Plot below represents the difference of predicted values from actual.



Mean Absolute Error : 1.3082

Mean Square Error : 28.1828