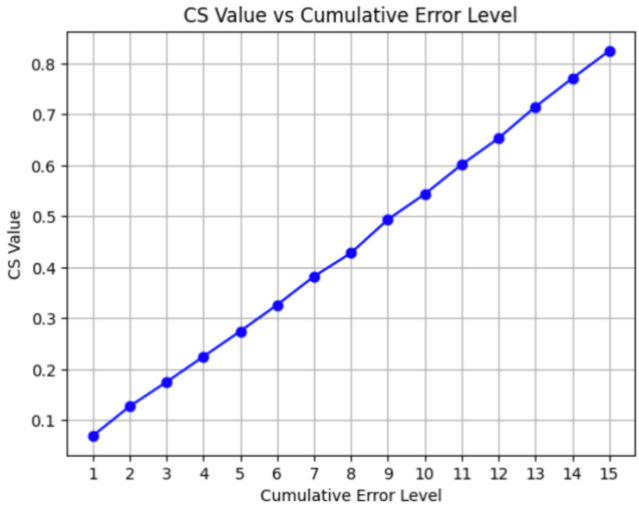
Name: Tianxiu Ma
Student number: 230277302
Assignment number: Assignment 3
module code: ECS797U/ECS797P

Task 3: Vary the cumulative error level from 1 to 15 and generate a plot of the CS value against the cumulative error level. What do you observe? [4 marks]



The observation results are as shown in the figure above:
The CS value increases with the increase in the error level. This is because, as the cumulative error level (i.e., the tolerated range of error) increases, the permissible range of error between the predicted age and the actual age also increases, thus increasing the number of samples considered to be accurately predicted, leading to a rise in the CS value.

Task 4: Compute the MAE and CS values (with cumulative error level of 5) for both partial least square regression model and the regression tree model. Compare with the previous method [4 marks]

```
linear regression:
linear regression Average MAE: 9.605030135804439
linear regression CS (Error <= 5): 0.2749003977533356
PLS Regression:
PLS Regression MAE: 6.843259635223232
PLS Regression CS (Error <= 5): 0.4581673306772908
Regression Tree:
Regression Tree MAE: 8.428286852589641
Regression Tree CS (Error <= 5): 0.5358565737051793
```

Partial Least Squares Regression is a method for extracting key data from a vast dataset, helping us to identify the most important information for making predictions

amidst large and complex data. Regression Trees make predictions, such as age, through a series of yes/no questions about features within the data. Each decision taken brings us closer to the final predicted outcome.

From the results provided, it can be observed that:
Partial Least Squares Regression performs best in terms of Mean Absolute Error (MAE), meaning that its predictions are closer to the actual ages on average. Additionally, its Cumulative Score (CS) indicates that over 45% of the predictions have an error of 5 years or less, although this is not the highest among the three methods.
The Regression Tree has the highest CS value, indicating that over 53% of its predictions have an error within 5 years. This suggests that the Regression Tree model may have better consistency in its predictions, even though its MAE is slightly higher than that of the PLS Regression.
Linear Regression performs the worst both in terms of MAE and CS values among these three methods.

Task 5: Compute the MAE and CS values (with cumulative error level of 5) for Support Vector Regression. Compare with the previous methods. How do you explain the results in the method comparison? Are the scores expected? [4 marks]

Support Vector Regression：

```
Support Vector Regression MAE: 8.525634734834059
Support Vector Regression CS (Error <= 5): 0.32270916334661354
```

The MAE of Support Vector Regression (SVR) is positioned between that of linear regression and Partial Least Squares Regression, and higher than that of the Regression Tree. This suggests that, for this particular task of age prediction, SVR's accuracy is better than the accuracy of simple linear regression but not as good as that of Regression Tree.
Regarding the CS value, SVR scores lower than both Partial Least Squares Regression and Regression Tree, indicating a lower accuracy for predictions within an error margin of 5 years or less.

The results of the method comparison are as follows:
Partial Least Squares Regression typically performs well when dealing with multiple correlated features and moderately-sized datasets, so its better performance in terms of MAE was expected in this context.
Regression Trees are known to be adept at handling non-linear problems. Their higher CS values reflect their capacity to deal with more complex data relationships than linear models.
SVR is a tool that uses known information (features) to predict specific numerical outcomes. It has some degree of tolerance for error, allowing for a margin of error in

predictions, and it can also help us handle very complex scenarios to find the best prediction method. In data with significant noise, SVR might not always show the best performance, especially without proper tuning of the kernel function and other hyperparameters.
The poor performance of linear regression might be due to the non-linear relationships between features and the age prediction problem, as well as the model's lack of capacity to capture complex relationships in the data.

Overall, the scores of the above models are within expectations.