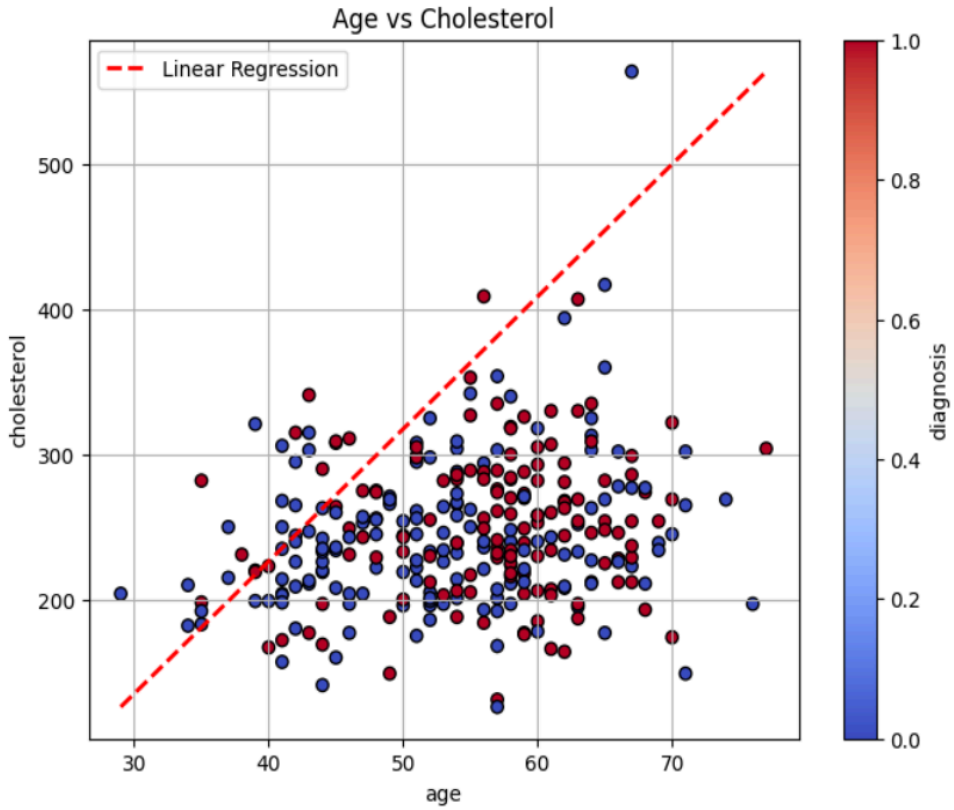
Name: Escalicas, Peter Allen, de Leon Bryan	**Date:** 3/5/24
Section: BsCPE32S3	**Professor:** Engr. Roman

PDF Report

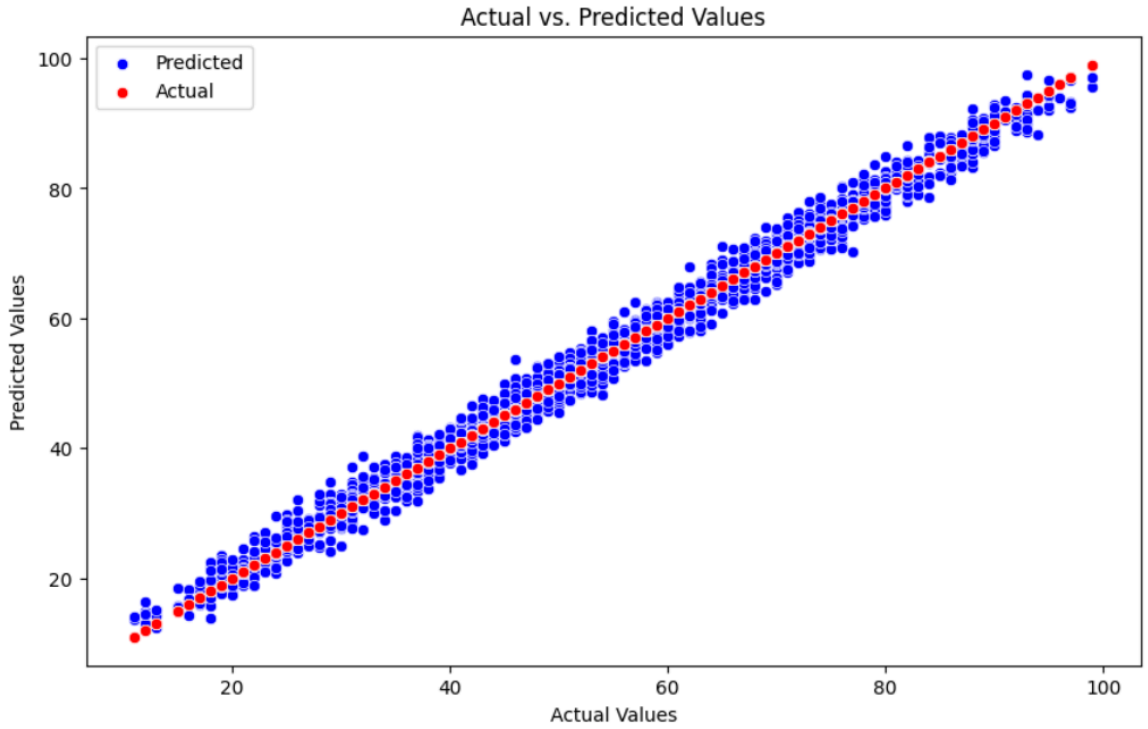
Single Linear Regression

- For our Single linear regression we want to know the relation of Age and Cholesterol in our chosen dataset. In this case we have used the variables Age and cholesterol, and for our target variable is the Diagnosis. In the plotting provided, we can see that a connection between age and cholesterol, with levels generally rising as people get older. The graph also reveals a peak around ages 50 to 60, where the most people have cholesterol levels between 200 and 300.



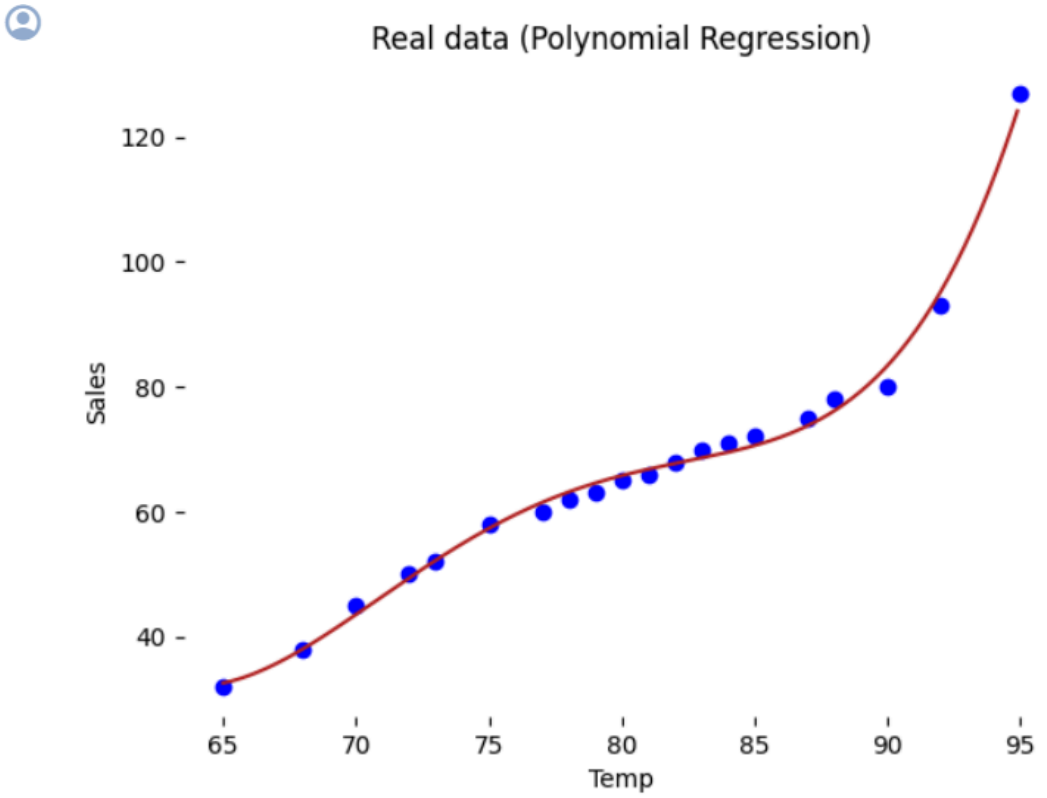
Multiple Linear Regression

- For our Multiple Linear Regression, we have set our target variable as the Performance Index in our dataset. But before we could train our model we have to slightly modify our dataset since we have a column data has a data of Female and Male. We have to turn that varchar datatype to a integer datatype in order for our model to work. After converting the datatype we were able now to train the model, and to our surprise it worked well since we can see at the plot that the prediction line is within the actual values of the dataset.



Polynomial Linear Regression

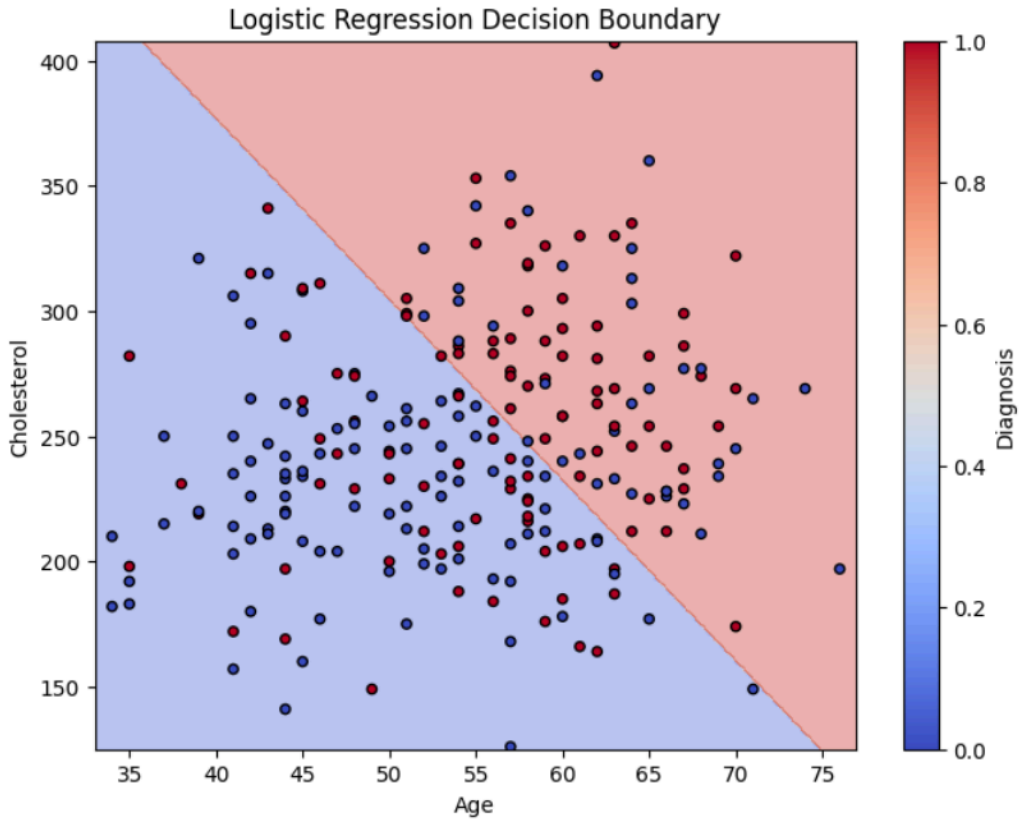
- For our Polynomial Linear regression we want to know the prediction of the sales as the Temperature rises. We can observed after a certain point in the Temperature the sales went up very dramatically, this is due to the extreme heat that the Ice cream sales skyrocketed. we can see that there is indeed a positive correlation between the temperature and sales, it suggest that people tend to buy more cold products amidst the hot season in order to alleviate the heat. Also the model can be predict the sales proportional to the temperature since we can see that the prediction line in not erratic.



Logistic Linear Regression

- A Logistic Linear Regression model tends to predict the likelihood of an event happening. In this case, age and cholesterol is used as a basis in order for the model to predict the likelihood of a person developing heart diseases, the decision boundary then separates the region where a positive diagnosis is more likely to happen.

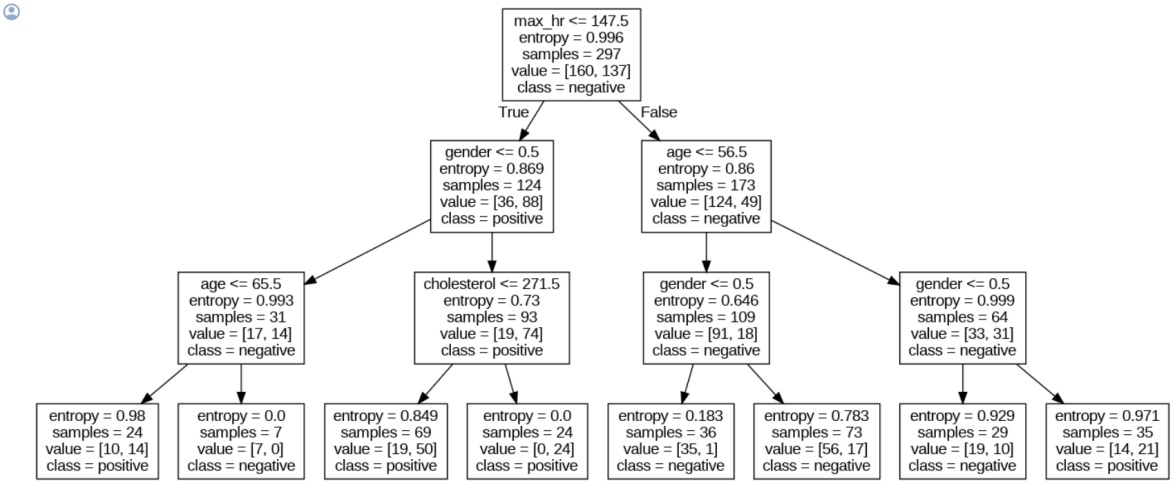
- Although we can see that there are alot of positive patients in the blue section of the decision boundary, the model instead tagged the region of age ranging from 60 and above as the positive boundary, as the positive patients are more concentrated here and there are few cases of negative patients.



Decision Tree

- In the decision tree model, the algorithm proceeds to select the feature that will best split the dataset into distinct groups, wherein at each point the tree splits into different branches. In the context of our data, the model weighs which features are more important for making accurate predictions for diagnosis, it does this by how well the features split the dataset into distinct groups.

- We can observe that in the cholesterol < 271.5 node, the model will proceed to split the data into two sets of groups, people exceeding the 271.5 cholesterol threshold and people with less than the threshold, both nodes exhibiting positive diagnosis class, meaning that cholesterol levels have a strong relationship with the likelihood of developing heart diseases. Gender on the other hand, after splitting the dataset into male and female, we can see that both nodes exhibit negative diagnosis, signifying that gender does not have a strong relationship to the likelihood.



Random Forest

- Random Forest utilizes an ensemble learning technique wherein multiple individual decision trees are generated and then merged in order to arrive at a final prediction. In this case, we can see that the classifier drew decision boundaries over the negative diagnosis plot points.

- In the generated tree figure, each node of the tree is trained on a random subset of our given dataset, once all nodes of the tree have been trained, their output is merged in order to arrive at a more accurate prediction.