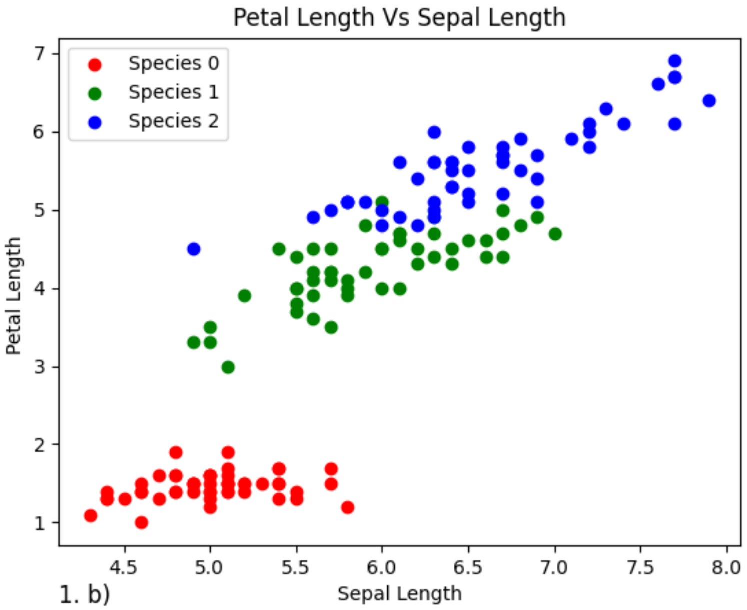
Take-Home Coding Assessment: L2 Assistant Data Scientist

Wilson Neira

1 Regression Modelling

a) 50 irises belong to species 0, 50 irises belong to species 1, and 50 irises belong to species 2.

b) Species 0 had about the smallest, species 1 had the second smallest, and species 2 had the largest sepal and petal length. There is a sepal and petal length positive relationship for species 1 and 2. There is a not so strong sepal and petal length relationship for when compared to species 1 and 2.



1. b)

d) My regression model has 3 coefficients which corresponds to the given predictors. The first coefficient 0.7110629145526267 represents the change in petal length for a one-unit increase in sepal length, holding all other predictor variables constant. Shown by the coefficient we can see that the relationship between the sepal length and petal length is positive, so when sepal length increases, petal length also increases. By the regression model R-squared value we notice that there exists a strong relationship between the predictors and petal length.

e) My regression model has 4 coefficients which corresponds to the given predictors. The first coefficient 0.7440377377222421 represents the change in petal length for a one-unit increase in sepal length, holding all other predictor variables constant. Shown by the coefficient we can see that the relationship between the sepal length and petal length is positive and greater than part d, so when sepal length increases, petal length also increases. By the regression model R-squared value 0.8626365429394165 we notice that it is greater than part d and there exists a strong relationship between the predictors and petal length.

2 Implementing an Edit-Distance Algorithm

2. 1) a. ”Kitten” and ”kitten” have a distance of 0.

2. 1) b. ”kitten” and ”KitTen” have a Hamming distance of 0.5.

1) c. ”Puppy” and ”POppy” have a distance of 1.5 (1 for the different letter, additional .5 for the different capitalization).

Test Outcome 1 Expected 1
Test Outcome 0.5 Expected 0.5
Test Outcome 2 Expected 2
Test Outcome 3.5 Expected 3.5

2) a) ”data Science” to ”Data Sciency” have a distance of 1.

2) b) ”organizing” to ”orGanising” have a Hamming distance of 0.5.

2) c) ”AGPRklafsdyweIllIIgEnXuTggzF” to ”AgpRkliFZdiweIllIIgENXUTygSF” have a distance of 6.5.

2) a) The Hamming distance algorithm would be applicable in social policy, the standard Hamming distance algorithm can be applied to analyze social media data to identify potential hate speech or discriminatory language. By representing social media posts or comments as binary strings or using one-hot encoding, the Hamming distance can be used to compare posts with known patterns of hate speech and identify potential instances of hate speech or discrimination. This can help policymakers to develop targeted interventions to reduce hate speech and promote inclusivity. For example, the Hamming distance can be used to compare social media posts with known patterns of hate speech and identify potential patterns of discriminatory language that may require further investigation or intervention.

3 Data Cleaning

a) The number of field descriptions reference a perspective that is not standard is 3455.

b) The average number of drawing descriptions per patent is 7.4416058394160585.