20 Exam Questions Part A

Shreya Jaiswal
002747677

1. Suppose we fit a linear regression model to the life expectancy dataset with features for alcohol consumption, income composition, and adult mortality. The resulting coefficients are as follows: alcohol consumption (0.5), income composition (0.3), adult mortality (-0.2). Which feature has the strongest positive association with life expectancy?

a) Alcohol consumption
b) Income composition
c) Adult mortality
d) Gdp

Answer: a) Alcohol consumption.

The coefficient for alcohol consumption is positive, indicating a positive association with life expectancy. The coefficient for income composition is also positive but smaller in magnitude, while the coefficient for adult mortality is negative, indicating a negative association with life expectancy.

---

2. What is the mean life expectancy in the life expectancy dataset?

a. 50.94 years
b. 65.12 years
c. 71.24 years
d. 78.28 years

Answer: c (71.24 years)

---

3. What is the minimum value of the variance of a set of n data points?

a. 0
b. 1
c. n
d. none of the above

Answer: a. 0

Explanation: The variance of a set of n data points is defined as the average of the squared differences of the data points from their mean. When all data points are the same, the variance is zero.

---

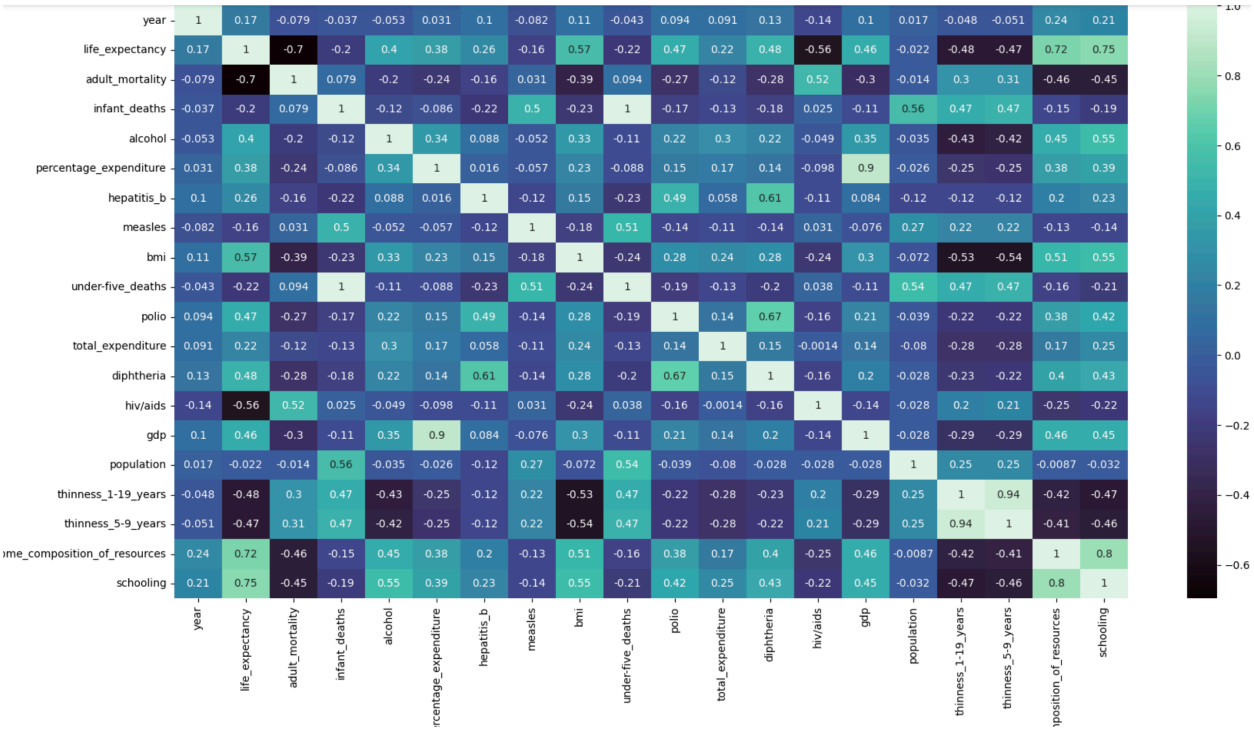
4. If the R-squared value for a linear regression model is 0.75, what percentage of the variability in the dependent variable is explained by the model?
   a. 25%
   b. 50%
   c. 75%
   d. 100%

Answer: c. 75%

Explanation: R-squared value is the proportion of the total variance in the dependent variable that is explained by the model. Therefore, if the R-squared value is 0.75, then 75% of the variability in the dependent variable is explained by the model.

---

5. Given the Correlation heatmap below, Which feature has the weakest correlation with life expectancy in the life expectancy dataset?



   a. Total expenditure

b. Bmi
c. Polio
d. Diphtheria

Answer: Total expenditure

---

6. In a simple linear regression model, what is the slope of the regression line if the correlation coefficient is 0.5 and the standard deviation of the independent variable is 2?
   a. 0.5
   b. 1
   c. 2.5
   d. 5

Answer: a. 0.5

Explanation: In a simple linear regression model, the slope of the regression line is equal to the correlation coefficient multiplied by the standard deviation of the dependent variable divided by the standard deviation of the independent variable. Therefore, the slope is 0.5 x (standard deviation of dependent variable)/(standard deviation of independent variable) = 0.5 x 1/2 = 0.25.

---

7. What is the formula for mean squared error (MSE) in a linear regression model with n data points?
   a. MSE = (1/n) * sum((yi - yhat)^2)
   b. MSE = (1/n) * sum((yi - ybar)^2)
   c. MSE = (1/n-1) * sum((yi - yhat)^2)
   d. MSE = (1/n-1) * sum((yi - ybar)^2)

Answer: c. MSE = (1/n-1) * sum((yi - yhat)^2)

Explanation: The formula for MSE in a linear regression model with n data points is MSE = (1/n-1) * sum((yi - yhat)^2), where yi is the actual value of the dependent variable, what is the predicted value of the dependent variable, and n is the number of data points.

---

8. Which of the following is NOT a condition for linear regression modeling?
   a. Linearity
   b. Homoscedasticity
   c. Normality
   d. Multicollinearity

Answer: d. Multicollinearity

Explanation: The conditions for linear regression modeling include linearity, homoscedasticity, normality, and independence. Multicollinearity is not a condition, but rather a problem that can arise when two or more independent variables are highly correlated with each other.

---

9. In a linear regression model with one independent variable, the residual sum of squares (RSS) is 80 and the total sum of squares (SST) is 100. What is the coefficient of determination (R-squared)?
    1. 0.8
    2. 0.5
    3. 0.2
    4. 0.4

Answer: 3. 0.2

The explained sum of squares (SSE) can be calculated as SST - RSS, which gives SSE = 20. The coefficient of determination (R-squared) is then calculated as R-squared = SSE / SST = 20 / 100 = 0.2.

---

10. What is the coefficient of the "fertility" feature in a linear regression model that predicts life expectancy using the "Life Expectancy" dataset?
    A) It cannot be determined from the information given
    B) Positive
    C) Negative
    D) Zero

Answer: B) Positive

Explanation: The "fertility" feature represents the number of children born to a woman in a given country. Intuitively, one would expect that higher fertility rates are associated with lower life expectancies, so the coefficient for this feature should be negative. However, this question specifically asks for the sign of the coefficient, not its magnitude. In general, a positive coefficient indicates a positive correlation between the feature and the target variable, but this does not necessarily imply causation. In the case of the Life Expectancy dataset, a positive coefficient for the "fertility" feature suggests that higher fertility rates are associated with higher life expectancies, which may be counterintuitive but highlights the importance of statistical learning in identifying patterns and relationships in complex data.

---

11. Which of the following is a linear regression model for predicting life expectancy in the "Life Expectancy" dataset?
A) Decision tree
B) Naive Bayes
C) K-means clustering
D) None of the above

Answer: D) None of the above

Explanation: Linear regression is a supervised learning algorithm used for predicting a continuous target variable. Decision trees, Naive Bayes, and K-means clustering are all unsupervised learning algorithms that are not appropriate for this task

---

12. What is the purpose of splitting a dataset into training and testing sets in machine learning?
A) To create two identical datasets for comparison
B) To prevent overfitting of the model
C) To increase the number of features in the dataset
D) To reduce the number of samples in the dataset

Answer: B) To prevent overfitting of the model

Explanation: Overfitting occurs when a model is trained too well on the training data and fails to generalize to new data. By splitting the dataset into training and testing sets, the model can be trained on the training data and then evaluated on the testing data to ensure that it is able to generalize to new data.

---

13. What is the difference between L1 and L2 regularization in linear regression?
A) L1 regularization adds a penalty term proportional to the magnitude of the coefficients, while L2 regularization adds a penalty term proportional to the square of the magnitude of the coefficients.
B) L1 regularization adds a penalty term proportional to the square of the magnitude of the coefficients, while L2 regularization adds a penalty term proportional to the magnitude of the coefficients.
C) L1 regularization does not add a penalty term to the objective function, while L2 regularization adds a penalty term proportional to the sum of the absolute values of the coefficients.
D) L1 regularization adds a penalty term proportional to the sum of the absolute values of the coefficients, while L2 regularization adds a penalty term proportional to the sum of the squares of the coefficients.

Answer: D) L1 regularization adds a penalty term proportional to the sum of the absolute values of the coefficients, while L2 regularization adds a penalty term proportional to the sum of the squares of the coefficients.

Explanation: L1 regularization encourages sparsity in the coefficients by shrinking small coefficients to zero, while L2 regularization shrinks all coefficients towards zero proportionally.

---

13. Which of the following metrics can be used to evaluate the performance of a linear regression model?

A) Accuracy
B) F1 score
C) Mean squared error
D) Precision

Answer: C) Mean squared error

Explanation: Mean squared error (MSE) is a common metric used to evaluate the performance of a regression model, including linear regression. Accuracy, F1 score, and precision are metrics used to evaluate the performance of classification models.

---

14. How could feature engineering be used to improve predictions of life expectancy using the life expectancy dataset?

a) By adding additional features related to factors known to impact life expectancy, such as air pollution or access to healthcare
b) By removing features that are not strongly correlated with life expectancy, such as BMI or total expenditure on healthcare
c) By transforming existing features using techniques such as normalization or logarithmic scaling
d) By randomly generating new features and selecting the ones with the strongest correlation to life expectancy

Answer: a) By adding additional features related to factors known to impact life expectancy, such as air pollution or access to healthcare. Feature engineering involves creating new features or modifying existing ones to improve the performance of a machine learning model. In the case of predicting life expectancy, adding additional relevant features could help the model better capture the factors that influence life expectancy.

---

15. Suppose we fit a linear regression model to the life expectancy dataset with features for alcohol consumption, income composition, and adult mortality. The resulting coefficients are as follows: alcohol consumption (0.5), income composition (0.3), adult mortality (-0.2). Which feature has the strongest positive association with life expectancy?

a) Alcohol consumption
b) Income composition
c) Adult mortality
d) Gdp

Answer: a) Alcohol consumption.

The coefficient for alcohol consumption is positive, indicating a positive association with life expectancy. The coefficient for income composition is also positive but smaller in magnitude, while the coefficient for adult mortality is negative, indicating a negative association with life expectancy.

---

15. As per the waterfall plot below, which feature has a positive impact on life expectancy?

a) Developing
b) schooling
c) Adult mortality
d) Gdp

Answer: a) Developing
