Machine Learning Assignment answers

1. A) Least Square Error
2. A) Linear regression is sensitive to outliers
3. B) Negative
4. B) Correlation
5. C) Low bias and high variance
6. B) Predictive modal
7. D) Regularization
8. D) SMOTE
9. A) TPR and FPR
10. B) False
11. C) Removing stop words
12. A) We don't have to choose the learning rate.
    B) It becomes slow when the number of features is very large.
    D) It does not make use of the dependent variable.

Subjective answers:

13. It is a technique used in machine learning and statistical modeling to prevent overfitting and reduce the complexity of a model. Regularization introduces a penalty term to the objective function being optimized during the training process. This penalty discourages the model from fitting the training data too closely and helps generalize better to unseen data. Overfitting takes place when a model learns not only the underlying patterns in the training data but also the noise and random fluctuations present in the data. It helps in achieving a balance between fitting the training data well and avoiding overfitting by penalizing overly complex models.
14. Several algorithms use regularization to prevent overfitting. Some are given below:
    - Lasso Regression(L1 Regularization)-This linear regression variant adds the sum of the absolute values of the coefficients as a penalty term to the loss function.
    - Ridge Regression(L2 Regularization)- This linear regression variant adds the sum of the squared values of the coefficients as a penalty term to the loss function.
    - Elastic Net-Combines both L1 and L2 regularization terms in the loss function, providing a balance between variable selection (L1) and coefficient shrinkage (L2).
    - Logistic regression with regularization-Logistic regression can also be regularized using L1 or L2 regularization to prevent overfitting.
15. The term "error" typically refers to the residuals or the difference between the predicted values and the actual values in the dataset. The linear regression equation models the relationship between the independent variables and the dependent variable as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots \beta_n X_n + \epsilon$$

The error term (ϵ) captures the unobserved factors or noise that affect the dependent variable but are not accounted for by the independent variables. Minimizing the sum of squared errors helps in finding the line that best represents the relationship between the variables in the given data