Machine Learning Worksheet

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 model
7. D) Regularization
8. D) SMOTE
9. A) TPR and FPR
10. B) False
11. B) Apply PCA to project high dimensional data
12. A) We don't have to choose the learning rate. B) It becomes slow when number of features is very large.
13. Regularization is a technique used in statistical modeling and machine learning to prevent overfitting, which occurs when a model learns the noise in the training data rather than the underlying pattern. By adding a penalty term to the loss function, regularization discourages overly complex models, leading to better generalization on unseen data.
    There are two common types of regularization:
    Lasso Regression (L1 Regularization): This adds a penalty equal to the absolute value of the coefficients. It can shrink some coefficients to zero, effectively performing variable selection.
    Ridge Regression (L2 Regularization): This adds a penalty equal to the square of the coefficients. It tends to keep all variables but reduces their impact,

preventing any one feature from dominating the model.
Regularization helps maintain a balance between fitting the training data well and keeping the model simple, thereby improving performance on new data.

14. Several algorithms incorporate regularization techniques to improve model performance and prevent overfitting. Here are some of the most commonly used algorithms:

1. **Lasso Regression (L1 Regularization)**: This algorithm adds a penalty equal to the absolute value of the coefficients. It can lead to sparse models by driving some coefficients to zero.
2. **Ridge Regression (L2 Regularization)**: This algorithm adds a penalty equal to the square of the coefficients. It reduces the impact of less important features without eliminating them entirely.
3. **Elastic Net**: This combines both L1 and L2 regularization. It is particularly useful when there are many correlated features, as it can select groups of correlated variables together.
4. **Regularized Logistic Regression**: Both Lasso and Ridge can be applied to logistic regression to handle binary classification problems while preventing overfitting.
5. **Support Vector Machines (SVM)**: SVMs can include regularization parameters that control the trade-off between maximizing the margin and minimizing classification error.
6. **Regularized Decision Trees (e.g., CART)**: Techniques like pruning can be seen as a form of regularization, helping to simplify the model.
7. **Neural Networks**: Regularization techniques like L1/L2 weight decay, dropout, and early stopping are commonly used to prevent overfitting in deep learning models.

These algorithms and techniques help ensure that models generalize well to new, unseen data while balancing complexity and performance.

15. In linear regression, the term "error" refers to the difference between the actual values of the dependent variable and the values predicted by the regression model. This discrepancy is crucial for understanding how well the model fits the data.

Key Aspects of Error in Linear Regression:

1. **Definition**:
    - The error for a specific observation can be defined as: Errori = yi − y^I where yi is the actual value and y^I is the predicted value from the regression model.
2. **Residuals**:
    - The errors are often referred to as "residuals" in the context of regression. Residuals are the differences between observed and predicted values across all observations in the dataset.
3. **Total Error**:
    - The overall error of the model can be quantified using various metrics, such as:
        - **Mean Squared Error (MSE)**
        - **Root Mean Squared Error (RMSE)**
4. **Interpretation**:
    - Smaller error values indicate a better fit of the model to the data. High error values suggest that the model is not capturing the relationship between the independent and dependent variables effectively.

5. **Assumptions**:
    - Linear regression assumes that the errors (residuals) are normally distributed, have constant variance (homoscedasticity), and are independent of one another.

Understanding and analyzing errors in linear regression is essential for improving model performance and ensuring that the predictions are as accurate as possible.