**PYTHON – WORKSHEET 1**

1. C
2. B.
3. A
4. A
5. D
6. C
7. A
8. C
9. A and C
10. A and B

**STATISTICS WORKSHEET-1**

1. A
2. A
3. B
4. C
5. C
6. B
7. B
8. A
9. C

Answer10:

The normal distribution describes a symmetrical plot of data around its mean value, where the width of the curve is defined by the standard deviation. It is visually depicted as the "bell curve."

Normal distribution is achieved/displayed when skewness is 0, kurtosis =3 and its shape is like a bell curve

Answer11: Missing value in our data

* We can replace missing values with mean, median or mode. We implement this when the data is missing at random and not related to the underlying pattern.
* Using Linear Regression model
* Using the average of the KNN values

Answer12: A/B testing

A/B testing, also known as split testing, is a method used in experimental design to compare two versions of a webpage, app, marketing campaign, or any other element with the goal of determining which version performs better.

The process is: Hypothesis> Variations> Randomization> Data Collection> Statistical Analysis> Conclusion.

Answer13:

Mean imputation means replacing the missing value with the mean of non-missing value of that feature. Its advantages are, it preserves Sample size, and it preserves relationships.

Answer14:

Linear Regression is a method used to model the relationship b/w a dependent and one or more independent variables by fitting a linear equation to observed data.

The basic formula is y=mx+c

Answer15:

Descriptive Statistics: This branch involves methods for summarizing and describing data, such as calculating measures of central tendency (mean, median, mode), measures of dispersion (range, variance, standard deviation), and graphical representations (histograms, box plots).

Inferential Statistics: Inferential statistics is concerned with making inferences or predictions about populations based on sample data. It includes techniques like hypothesis testing, confidence intervals, and regression analysis.

Predictive statistics: also known as predictive analytics or predictive modeling, is a branch of statistics and data analysis that involves using historical data to make predictions about future events or outcomes.

**MACHINE LEARNING**

1. D
2. A
3. B
4. B
5. C
6. B
7. D
8. D
9. A
10. B
11. A
12. D
13. Explain the term regularization?

Ans: Regularization is a technique used in machine learning and statistical modeling to prevent overfitting and improve the generalization performance of a model. Overfitting occurs when a model learns to fit the training data too closely, capturing noise and idiosyncrasies in the data that don't generalize well to new, unseen data. Regularization introduces a penalty term to the model's loss function, discouraging the model from learning overly complex relationships that might not hold in the broader population.

There are different types of regularization techniques, commonly used in linear regression and related algorithms:

1. L1 Regularization (Lasso): In L1 regularization, a penalty term proportional to the absolute values of the model's coefficients is added to the loss function. This encourages the model to have sparse coefficient values, effectively performing feature selection by forcing some coefficients to become exactly zero.

2. L2 Regularization (Ridge): L2 regularization adds a penalty term proportional to the squared magnitudes of the model's coefficients to the loss function. This technique encourages smaller but non-zero coefficient values, helping to avoid extreme values and reducing the impact of less important features.

3. Elastic Net Regularization: Elastic Net combines both L1 and L2 regularization, striking a balance between feature selection (Lasso) and coefficient shrinkage (Ridge).

Regularization helps to prevent overfitting by adding a bias towards simpler models. It's important to choose the right strength of regularization (controlled by a hyperparameter) to strike a balance between fitting the training data well and generalizing to new data effectively. Regularization is especially useful when working with high-dimensional datasets or when the number of features is large compared to the number of observations.

1. Which particular algorithms are used for regularization?

Ans: Several machine learning algorithms utilize regularization techniques to improve model generalization and prevent overfitting. Some of the commonly used algorithms that incorporate regularization are:

1. Linear Regression with Lasso (L1) and Ridge (L2) Regularization: Lasso Regression and Ridge Regression are variants of linear regression that include L1 and L2 regularization, respectively. They modify the loss function to penalize the magnitudes of the coefficients, leading to feature selection (Lasso) and coefficient shrinkage (Ridge).
2. Logistic Regression with L1 and L2 Regularization: Similar to linear regression, logistic regression can be regularized using L1 or L2 regularization. Regularized logistic regression can help improve classification performance by preventing overfitting.

3. Support Vector Machines (SVM) with Regularization: SVMs can be regularized using techniques like the C parameter, which controls the trade-off between maximizing the margin and minimizing the classification error. Higher values of C can lead to less regularization, while lower values encourage a wider margin and more regularization.

4. Decision Trees with Pruning: Decision trees can be regularized through pruning, where branches that do not contribute significantly to the model's accuracy are pruned (removed). Pruning prevents the tree from growing too deep and fitting noise in the training data.

5. Regularized Random Forests: Random Forests, an ensemble learning method, can benefit from regularization by using techniques like limiting tree depth, adjusting the number of features considered for splitting, or applying bagging with bootstrapped samples.

Regularization techniques can be applied to a variety of algorithms to enhance their ability to generalize from training data to new, unseen data. The choice of regularization technique and its hyperparameters depends on the specific problem and dataset characteristics.

1. Explain the term error present in linear regression equation?

Ans: In linear regression, the term "error" refers to the difference between the observed (actual) values of the dependent variable and the predicted values generated by the linear regression model. These errors, also known as residuals, represent the variability in the data that is not explained by the linear relationship between the independent variables and the dependent variable.