WORKSHEET 1

PYTHON

1. C) %
2. B) 0
3. C) 24
4. A) 2
5. D) 6
6. B) It encloses the lines of code which will be executed if any error occurs while executing the lines of code in the try block.
7. A) It is used to raise an exception.
8. C) in defining a generator
9. C) abc2
10. A)yield and B)raise

STATISTICS

1. d. The probability of rejecting H0 when H1 is true
2. b. null hypothesis
3. d. Type 1 error
4. b. the t distribution with n-1 degrees of freedom
5. a. accepting H0 when it is false
6. d. two-tailed test
7. b. the probability of committing a Type 1 error
8. a. the probability of committing a Type 2 error
9. b.
10. c. the level of significance
11. a. level of significance
12. d. All of the Above
13. ANOVA (ANalysis Of Variance) is used to compare the differences of means among more than 2 groups. It does this by looking at variation in the data and where that variation is found.
14. ANOVA can be used only when:
    - The sample have a normal distribution
    - The samples are selected at random and should be independent of one another
    - All groups have equal standard deviations.

15.

BASIS FOR COMPARISON	ONE WAY ANOVA	TWO WAY ANOVA
Meaning	It is a hypothesis test, used to test the equality of three of more population means simultaneously using variance.	It is a statistical technique wherein, the interaction between factors, influencing variable can be studied.
Independent Variable	One	Two
Compares	Three or more levels of one factor.	Effect of multiple levels of two factors.
Number of Observation	Need not to be same in each group.	Need to be equal in each group.
Design of experiments	Need to satisfy only two principles.	All three principles need to be satisfied.

MACHINE LEARNING

1. B)
2. D)
3. D)
4. C)
5. D)
6. C)
7. C)
8. A) & B)
9. C) & D)
10. A)
11. When the number of categorical data in the dataset is very large then we should avoid using One Hot Encoding that lead to high memory consumption. In this case Binary encoding will be used.

12. Techniques to be used to balance the imbalanced Dataset:

    1. Choose Proper Evaluation Metric:
    For an imbalanced class dataset F1 score is a more appropriate metric. It is the harmonic mean of precision and recall and the expression is –
    F1 = 2*[(precision* recall) / (precision + recall)]

    2. Resampling (Oversampling and Undersampling): When we are using an imbalanced dataset, we can oversample the minority class using replacement. This technique is called oversampling. Similarly, we can randomly delete rows from the majority class to match them with the minority class which is called under sampling. After sampling the data we can get a balanced dataset for both majority and minority classes.

    3. SMOTE: Synthetic Minority Oversampling Technique or SMOTE is another technique to oversample the minority class. Simply adding duplicate records of minority class often don't add any new information to the model. In SMOTE new instances are synthesized from the existing data. If we explain it in simple words, SMOTE looks into minority class instances and use *k* nearest neighbor to select a random nearest neighbor, and a synthetic instance is created randomly in feature space.

13.

The main difference between SMOTE and ADASYN is that ADASYN uses a density distribution to decide the number of samples.

SMOTE: It find the n-nearest neighbors in the minority class for each of the samples in the class. Then it draws a line between the neighbors and generates random points on the lines.

ADASYN: It is a better version of SMOTE. It is generating minority data samples according to density distribution using K nearest neighbors.

14.

GridSearchCV is a library function of sklearn model selection package. It helps to loop through predefined hyperparameters and fit the estimator on the training set. In the end, we can select the best parameters from the listed hyperparameters.

We can use GridSearchCV on large dataset but it is time consuming. We can use RandomSearchCV, which uses random hyperparameter values to pick the best hyperparameter.

15. Evaluation metrics:

- Mean Absolute Error (MAE): It is calculated by taking the absolute difference between the predicted values and the actual values and averaging it across the dataset.
- Mean Squared Error (MSE): it is L2 loss; we calculate the error by squaring the difference between the predicted value and actual value and averaging it across the dataset.
- Root Mean Squared Error (RMSE): It measures the average magnitude of the errors and is concerned with the deviations from the actual value.