Assignment No. 1

Course Code: ECAP792 **Registration Number:** 322201297

Instructions:

a. Attempt all questions given below in your own handwriting. Assignment in typed format will not be considered for evaluation.

b. The student has to complete the assignment in the allocated pages only. Any other page in case utilized shall not be considered.

Q1.

a. How should missing values be handled in a dataset?

b. What distinguishes overfitting from underfitting?

c. How should outliers in a dataset be handled?

[10 Marks] [CO4, U04L1]

(a). Mising values in a dataset can be handled in several ways, one common methods is imputation, where missing values are filled in using stastical techniques such as mean, median, or mode imputation. Another approach is to replace missing values with zeroes, especially in cases where the missing values depends on the nature of the data and the analysis requirements.

(b) Overfitting and underfitting are two common issues in machine learning models. Overfitting occurs when a model is overly complex, capturing noise in the training data and leading to high variance and low bias. This results in the model performing well on training data but poorly on unseen test data. On other hand, underfitting happens when a model is too simple, lacking the capacity to capture the underlying patterns in the data. This leads to high bias and low variance, causing poor performance on both training and test data.

(c) Outless in a dataset can be handled through various techniques. Probabilities and statistical modeling involves identifying outliers based on the distribution of the data. Z-score analysis significantly far away. Proximity-based models consider the distance of data points from their neighbors to detect outliers. Linear regression models can be used to identify outliers based on the residuals of the model. High-dimensional outlier detection methods are suitable for datasets with many features, where traditional methods may not be effective. Removing outliers is crucial to ensure that the model is not skewed by these extreme values and provides more accurate predictions.

Signature of the Student__________

Note:-

CO: is the Course Outcome as per your course syllabus.

L1-L6: Learning level objectives as per Revised Bloom Taxonomy (RBT).

Assignment No. 1

Course Code: ECAP792 **Registration Number:** 322201297

Instructions:

a. Attempt all questions given below in your own handwriting. Assignment in typed format will not be considered for evaluation.

b. The student has to complete the assignment in the allocated pages only. Any other page in case utilized shall not be considered.

Q2. What is hypothesis testing. Explain difference between type1 and type2 error and also explain role of level of significance.

[10 Marks] [CO2, L2]

Hypothesis Testing

Hypothesis testing is a fundamental statistical method used to assess the validity of hypothesis regarding a population parameter based on sample data. Compares observed data with expected outcomes under the null hypothesis.

Type I Error	Type II Error
• Occurs when the null hypothesis is wrongly rejected, leading to a false positive. • Represents the probability of incorrectly rejecting a true null hypothesis.	• Happens when the null hypothesis is not rejected when it is false, resulting in false negative • Reflects the probability of failing to reject a false null hypothesis.

Level of Significance:

- Denoted by alpha. it sets the threshold for rejecting the null hypothesis
- Determines the probability of Commiting a Type I error.

Signature of the Student

Note:-

CO: is the Course Outcome as per your course syllabus.

L1-L6: Learning level objectives as per Revised Bloom Taxonomy (RBT).