DATE- 17/02/2023

PYTHON WORKSHEET-1

QNO1- Which of the operator is used to calculate the remainder in a division?

Ans- % operator.

QNO2- In python 2//3 is equal to:

Ans- 0

QNO3- in python, 6<<2

Ans- 24

QNO4- In python, 6&2 will give which of the following as output?

Ans- 2

QNO5-In python, 6|2 will give which of the following as output?

Ans- 6

QNO6-What does the finally keyword denotes in python?

Ans-the finally block will be executed no matter if the try block raises an error or not.(option c)

QN0-7- What does raise keyword is used for in python?

Ans- It is used to raise an exception.(Option A)

QNO 8- Which of the following is a common use case of yield keyword in python?

Ans**- In defining the generator (option c)**

QNO 9- Which of the following are the valid variable names?

Ans- **_abc & abc2 (option A & C)**

QN0-10- Which of the following are the keywords in python?

Ans- **yield & raise(option A & B)**

STATISTICS WORK SHEET

Q1. In hypothesis testing, type II error is represented by β and the power of the test is 1−β then β is:

Ans:- The probability of rejecting H0 when H1 is true(opt A)

Q-2 In hypothesis testing, the hypothesis which is tentatively assumed to be true is called the.

Ans:- Null hypothesis (opt-B)

Q-3 When the null hypothesis has been true, but the sample information has resulted in the rejection of the null, a _____ has been made.

Ans:- Type 1 error

Q-4 .For finding the p-value when the population standard deviation is unknown, if it is reasonable to assume that the population is normal, we use

Ans:- the t distribution with n + 1 degrees of freedom(opt-c)

Q-5 A Type II error is the error of.

Ans:- rejecting Ho when it is false (opt-c)

Q-6 A hypothesis test in which rejection of the null hypothesis occurs for values of the point estimator in either tail of the sampling distribution is called?

Ans:- a two-tailed test (opt-d)

Q-7 In hypothesis testing, the level of significance is

Ans:- the probability of committing a Type I error (opt-b)

Q-8 In hypothesis testing, b is?

Ans:- the probability of committing a Type II error (opt-a)

Q9:- . When testing the following hypotheses at an α level of significance H0: p = 0.7 H1: p > 0.7 The null hypothesis will be rejected if the test statistic Z is

Ans: Z>z α (option A)

Q-10:- Which of the following does not need to be known in order to compute the P-value? a. knowledge of whether the test is one-tailed or two-tail

Ans:- knowledge of whether the test is one tailed or two tail(option A)

Q-11:- The maximum probability of a Type I error that the decision maker will tolerate is called the?

Ans:- level of significance(option A)

Q-12:- For t distribution, increasing the sample size, the effect will be on?

Ans:- The t ratio (option-b)

Q-13: What is ANOVA in SPSS?

Ans: ANOVA (Analysis of Variance) in SPSS is a statistical method used to analyze the differences among means of two or more groups or treatments. ANOVA allows us to test whether there are any significant differences in the means of the groups being compared, and helps us determine which groups are significantly different from each other.

SPSS (Statistical Package for the Social Sciences) is a software program commonly used in data analysis, and it includes tools for performing ANOVA. In SPSS, ANOVA can be performed using the "General Linear Model" option, which includes various types of ANOVA such as one-way ANOVA, factorial ANOVA, and repeated measures ANOVA.

To perform ANOVA in SPSS, we need to enter the data into the program and specify the groups or treatments to be compared. We can then use the ANOVA procedure to obtain the F-value, which indicates whether there are significant differences among the means of the groups being compared. If the F-value is significant, we can use post-hoc tests to determine which groups are significantly different from each other.

Overall, ANOVA in SPSS is a powerful tool for analyzing differences among groups and treatments, and can be used in a variety of research fields including social sciences, health sciences, and engineering.

Q-14:- what are the Assumptions of ANOVA?

Ans:- The most important assumptions of ANOVA are:

1. Independence: The observations within each group are independent of each other.
2. Normality: The population distributions are normal or approximately normal.

3. Homogeneity of variances: The variances of the populations are equal or approximately equal.
4. Random Sampling: The observations are randomly selected from the population.
5. Interval or ratio data: The dependent variable is measured on an interval or ratio scale.

Violations of these assumptions can affect the accuracy and reliability of the ANOVA results.

Q-15:- What is the difference between one way anova and two way anova?

Ans:- One-way ANOVA and two-way ANOVA are both statistical methods used to compare means across two or more groups or treatments. The main difference between the two methods is the number of independent variables being considered.

One-way ANOVA, as the name suggests, involves only one independent variable (factor). It is used to test for differences in means among three or more groups on a single dependent variable. For example, we might use one-way ANOVA to test whether there are differences in the mean test scores among students who were taught using three different teaching methods.

Two-way ANOVA involves two independent variables (factors) and is used to test for differences in means among groups defined by the combination of the two factors. For example, we might use two-way ANOVA to test whether there are differences in the mean test scores among students who were taught using two different teaching methods and were given different amounts of homework.

In summary, the main difference between one-way ANOVA and two-way ANOVA is that one-way ANOVA involves a single independent variable, while two-way ANOVA involves two independent variables. The choice between these two methods depends on the research question and the variables being considered. If there is only one variable of interest, one-way ANOVA may be appropriate, while if there are two variables of interest, two-way ANOVA may be more appropriate.

Machine Learning sheet

Q-1: What is the advantage of hierarchical clustering over K-means clustering?

Ans:- In hierarchical clustering you don't need to assign number of clusters in beginning(option B)

Q-2: Which of the following hyper parameter(s), when increased may cause random forest to over fit the data?

Ans: max_depth(option A)

Q-3: Which of the following is the least preferable resampling method in handling imbalance datasets?

Ans:- SMOTE (option A)

Q-4:- Which of the following statements is/are true about "Type-1" and "Type-2" errors?

1. Type1 is known as false positive and Type2 is known as false negative
2. Type1 is known as false negative and Type2 is known as false positive
3. Type1 error occurs when we reject a null hypothesis when it is actually true.

Ans:- 2&3 (option D)

Q-5:- Arrange the steps of k-means algorithm in the order in which they occur:

1. Randomly selecting the cluster centroids

2. Updating the cluster centroids iteratively

3. Assigning the cluster points to their nearest centre.

Ans:- 1-3-2(option- D)

Q-6:- Which of the following algorithms is not advisable to use when you have limited CPU resources and time, and when the data set is relatively large?

Ans:- Support Vector Machine(SVM) (option B)

Q-7: What is the main difference between CART (Classification and Regression Trees) and CHAID (Chi Square Automatic Interaction Detection) Trees?

Ans: None of above(option- D)

Q-8: In Ridge and Lasso regularization if you take a large value of regularization constant(lambda), which of the following things may occur?

Ans:- A) Ridge will lead to some of the coefficients to be very close to 0.

B) Lasso will cause some of the coefficients to become 0. (option A and B)

Q-9: Which of the following methods can be used to treat two multi-collinear features?

Ans:- remove only one of the features(option-B)

Q-10: After using linear regression, we find that the bias is very low, while the variance is very high. What are the possible reasons for this?

Ans:- Overfitting (option A)

Q-11: . In which situation One-hot encoding must be avoided? Which encoding technique can be used in such a case?

Ans:- One-hot encoding should be avoided when dealing with high cardinality categorical variables, i.e., when the number of unique categories is very large. This can result in a high number of features after encoding, which can lead to computational and memory issues, as well as overfitting. In such cases, an alternative encoding technique such as target encoding or frequency encoding can be used, which replace each category with a single numerical value based on its relationship with the target variable or its frequency in the dataset, respectively. These techniques can help reduce the number of features and provide more meaningful representations of categorical variables in certain scenarios.

Q-12:- In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly

Ans:- Data imbalance occurs when the number of instances in one class is significantly higher than the number of instances in another class in a classification problem. In such cases, traditional machine learning algorithms may produce biased models that prioritize the majority class, resulting in poor performance for the minority class. To address this issue, the following techniques can be used to balance the dataset:

1. **Undersampling:** In undersampling, a random subset of the majority class is selected, and the size of the dataset is reduced to match the size of the minority class. However, undersampling may result in loss of information, and it may not be the best approach if the dataset is small.
2. **Oversampling:** In oversampling, the minority class is oversampled by creating synthetic examples by using techniques such as SMOTE (Synthetic Minority Over-sampling Technique). SMOTE generates synthetic examples by interpolating between existing examples in the

minority class. Oversampling can improve the performance of models, but it may also lead to overfitting.

3. **Class weighting:** In this technique, a weight is assigned to each class, depending on the proportion of instances in each class. The weight of the minority class is increased, which gives more importance to the minority class instances in the training process.

4. **Ensemble methods:** Ensemble methods such as Random Forest and Boosting can also be used to handle class imbalance problems. These methods combine multiple classifiers to improve the overall performance, and they can handle class imbalance by assigning more weight to the minority class during training.

It is important to note that the choice of the technique depends on the characteristics of the dataset and the problem at hand. A combination of these techniques can also be used to balance the dataset and improve the performance of the models.

Q-13: What is the difference between SMOTE and ADASYN sampling techniques?

Ans:- SMOTE (Synthetic Minority Over-sampling Technique) and ADASYN (Adaptive Synthetic Sampling) are both oversampling techniques used to balance imbalanced datasets in classification problems. However, there are some differences between the two techniques:

1. Approach: SMOTE generates synthetic samples for the minority class by interpolating between neighboring samples, whereas ADASYN generates synthetic samples by adapting the weights of the different samples based on their level of difficulty in learning.

2. Focus: SMOTE focuses on increasing the density of the minority class by generating synthetic samples, whereas ADASYN focuses on the minority samples that are harder to learn by generating more synthetic samples for them.

3. Sample generation: In SMOTE, synthetic samples are generated at a fixed rate between the nearest neighbors, whereas in ADASYN, the number of synthetic samples generated for each sample is proportional to the level of difficulty.

4. Noise: SMOTE generates synthetic samples by interpolation between the nearest neighbors, which may lead to noise in the dataset, whereas ADASYN adapts the weights of the samples, which reduces the noise generated in the dataset.

In summary, both SMOTE and ADASYN are effective techniques to balance imbalanced datasets, but their approach and focus on the minority samples differ. ADASYN is more effective in generating synthetic samples for harder to learn minority samples, while SMOTE is effective in generating synthetic samples for minority samples that are close to each other.

Q-14:- What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why or why not?

Ans:- The purpose of using GridSearchCV is to exhaustively search for the optimal hyperparameters for a given machine learning algorithm. It involves specifying a range of values for each hyperparameter and then evaluating the model's performance for each

combination of hyperparameters using cross-validation. GridSearchCV then returns the combination of hyperparameters that result in the best performance metric.

GridSearchCV can be useful in finding the best hyperparameters for a given algorithm, but it can be computationally expensive and time-consuming, especially for large datasets. In such cases, randomized search or Bayesian optimization may be preferred as they can sample a smaller subset of hyperparameters and converge to an optimal solution more quickly.

However, if the dataset is relatively small and the number of hyperparameters to be tuned is limited, GridSearchCV can still be a useful tool for finding the optimal hyperparameters.

Q-15:- List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief

Ans: Here are some common evaluation metrics used to evaluate a regression model:

1. Mean Squared Error (MSE): This metric measures the average squared difference between the predicted and actual values. It is calculated by taking the average of the squared differences between the predicted and actual values.
2. Root Mean Squared Error (RMSE): This is the square root of the MSE and is a popular metric for regression problems. It has the same unit as the dependent variable, making it easy to interpret.
3. R-squared (R2): This metric represents the proportion of variance in the dependent variable that can be explained by the independent variables in the model. It ranges from 0 to 1, with a higher value indicating a better fit.
4. Mean Absolute Error (MAE): This metric measures the average absolute difference between the predicted and actual values. It is calculated by taking the average of the absolute differences between the predicted and actual values.
5. Coefficient of Determination (COD): It is also known as R-squared (R2), which measures the proportion of variance explained by the model compared to the total variance in the data.
6. Mean Absolute Percentage Error (MAPE): This metric measures the percentage difference between the predicted and actual values. It is calculated by taking the average of the absolute percentage differences between the predicted and actual values.

The choice of evaluation metric depends on the problem statement and the domain. It is advisable to use multiple evaluation metrics to have a better understanding of the model's performance. GridSearchCV can be used to tune the hyperparameters of the model, but it may not be suitable for large datasets due to its computational cost.