❖ Worksheet set 1 - Machine Learning Assignment (Answers)

Q1 – B

Q2 – A

Q3 – C

Q4 – C

Q5- D

Q6- B

Q7- B

Q8- A&B

Q9- C & D

Q10- A

Q11 - 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 avoid when there is large number of categorical variable in the dataset which contains lots of unique values
The usage of encoding techniques is totally dependent on what kind of problem we are solving and what kind of data set we are dealing with. There are other encoding technique we can used like Label Encoder, frequency encoding and target encoding

Q12 - In case of data imbalance problem in classification, what techniques can be used to balance the dataset? Explain them briefly.

ANS - data set imbalance is the common problem in the classification In which the number of data of one variable is higher than other variable. To avoid this issue several balancing techniques can be used here are some of them:

1. Under sampling: This technique involves removing some of the instances from the majority class to balance the dataset. Under sampling is a simple technique that can be effective when the dataset is relatively large and the majority class has a significant number of instances.
2. Oversampling: This technique involves duplicating some of the instances from the minority class to balance the dataset. Oversampling can be effective when the dataset is relatively small, and the minority class has few instances.
3. Synthetic Sampling: Synthetic sampling involves generating synthetic instances of the minority class to balance the dataset. The most used synthetic sampling techniques are SMOTE and ADASYN. These techniques generate synthetic instances by interpolating between existing instances of the minority class.
4. Class Weighting: This technique involves assigning weights to each class based on the frequency of instances. Class weighting can be effective when the class imbalance is not too severe, and the number of instances is relatively large.
5. Ensemble methods: Ensemble methods involve combining multiple models trained on different subsets of the dataset. These methods can be effective in handling class imbalance, as they can assign more weight to the minority class and prevent the model from being biased towards the majority class.

13. What is the difference between SMOTE and ADASYN sampling techniques?

ANS -

1 – SMOTE is generate synthetic sample by focusing on minority class samples while the ADASYN is focusing on weighted distribution of synthetic sample with more reliable on harder to learn examples

2 - The primary distinction between ADASYN and SMOTE is that the former uses a density distribution as a criterion to automatically determine the number of synthetic samples that need to be generated for each minority sample by adapting the weights of the various minority samples to account for the skewed distribution

Q14 - What is the purpose of using GridSearchCV? Is it preferable to use in case of large datasets? Why or why not?

ANS - A technique for fine-tuning hyperparameters in machine learning models is called GridSearchCV. It enables a thorough search over a specified hyperparameter grid, improving the performance of the model. It is helpful for selecting the hyperparameters that work best together to improve model performance.It is possible to utilise GridSearchCV with both small and large datasets. The exhaustive search over a broad hyperparameter grid, however, may require more time when working with huge datasets, increasing the computational cost. Other hyperparameter optimization methods, such RandomizedSearchCV or Bayesian optimization, may be better suitable in such circumstances.

Q15 - List down some of the evaluation metric used to evaluate a regression model. Explain each of them in brief.

1. **Mean Squared Error (MSE)**: MSE measures the average of the squared differences between the predicted and actual values. It penalizes large errors more severely than small errors.
2. **Root Mean Squared Error (RMSE)**: RMSE is the square root of the MSE, and it is a measure of the average magnitude of the errors. It is expressed in the same units as the target variable.
3. **Mean Absolute Error (MAE)**: MAE measures the average absolute difference between the predicted and actual values. It is less sensitive to outliers than MSE.
4. **R-squared (R^2)**: R-squared is a statistical measure that represents the proportion of the variance in the target variable that is explained by the model. It ranges from 0 to 1, with 1 indicating a perfect fit.

These metrics can help assess the accuracy of generalization capabilities of regression models.It is important to choose the appropriate metric based on the specific problem and characteristics of the target variable

❖ Worksheet 1 Python Assignment Answers

Q1 – C

Q2 - A

Q3 - C

Q4 – A

Q5 – D

Q6 – C

Q7 – A

Q8 – C

Q9 – A&C

Q10- A&B

Question 11 to 15 is on Jupyter notebook

Worksheet 1 – Statistics

Q1 – c

Q2 - b

Q3- d

Q4- b

Q5 – a

Q6- d

Q7 – b

Q8 – a

Q9 – c

Q10 – c

Q11 – a

Q12 – c

Q13. What is Anova in SPSS?

Ans. In SPSS, ANOVA stands for Analysis of Variance, which is a statistical technique used to analyze the differences between two or more groups/variables/columns . ANOVA can be used to compare means across multiple groups and determine whether there are significant differences between the groups.

To conduct an ANOVA in SPSS, you first need to have your data organized into groups or categories. Once you have your data organized, you can use the ANOVA tool in SPSS to analyze the differences between the groups. SPSS provides several different ANOVA tools, including one-way ANOVA, two-way ANOVA, and mixed-design ANOVA.

The results of an ANOVA analysis in SPSS will provide information about the mean differences between the groups, as well as information about the significance of those differences. ANOVA results in SPSS can be displayed in a variety of ways, including tables and graphs, and can be used to inform further statistical analysis or decision-making.

14. What are the assumptions of Anova?

Ans. The assumptions of ANOVA (Analysis of Variance) include:

1. Normality: The data should be normally distributed within each group.
2. Homogeneity of variance: The variance of the dependent variable should be the same across all groups.
3. Independence: The observations within each group should be independent.

If these assumptions are not met, the results of the ANOVA may not be valid. Violations of these assumptions can lead to incorrect conclusions and can decrease the power of the test.

15. What is the difference between one way Anova and two way Anova?

Ans. The main difference between one-way ANOVA and two-way ANOVA is the number of independent variables or factors involved in the analysis.

In one-way ANOVA, there is only one independent variable or factor, which has three or more groups. In two-way ANOVA, there are two independent variables or factors that can affect the dependent variable.

The interaction effect between the two factors is also a key consideration in two-way ANOVA. The interaction effect refers to the combined effect of the two factors on the dependent variable, and it can be either additive or multiplicative.

In summary, while one-way ANOVA compares the means of three or more groups based on a single independent variable, two-way ANOVA compares the means of groups based on two independent variables and investigates the interaction effect between them.