**DATA SCIENCE TEST**

**MCQ**

**Question Answer**

1. **A, B, C**
2. **B**
3. **D**
4. **A**
5. **A, B**
6. **D**
7. **C**
8. **A**
9. **B**
10. **A**
11. **D**
12. **C**
13. **A, B, C**
14. **A, C**
15. **A**
16. **C**
17. **A**
18. **A**
19. **A**
20. **B, D**
21. **B, D**
22. **B**
23. **A**
24. **D**
25. **B**
26. **A**
27. **A**
28. **A**
29. **B**
30. **C**

**SCENARIO-BASED QUESTIONS**

**Q**

1. **To handle missing values in a dataset we can consider several techniques such as –**

* **Removing rows with missing values, if there are very few missing values.**
* **Removing missing values using mean, median, or mode of the columns.**
* **If all missing values is approax one, two percent we can remove all missing value along column.**
* **If missing values is very less we can ignore also.**

1. **For the prediction we can use algorithms like logistic regression, random forest, support vector machines…**

* **Logistic regression suitable for binary classification problems.**
* **Random forest can capture complex relationships between customer behaviours.**
* **SVM can used for separating different groups and are capable of capturing linear or non-linear.**
* **We can use metrics like accuracy, precision, f1 score, etc..**

1. **We use clustering algorithms like K-Means cluster, hierarchical cluster.**

* **Using k means we can train the data in different number of groups.**
* **Also we can focused on specific target customer according to their purchasing power.**
* **Which can improve customer retention and profitability.**

1. **For handling categorical variables in model, we can use techniques like one hot encoding, label encoding.**

* **One hot encoding creates binary columns for each category.**
* **Label encoding assigns unique integers to categories.**

**Concept Based Questions**

**Q**

1. **Multicollinearity is a phenomenon in regression analysis where two or more independent variables in a model.**

**To handle multicollinearity..**

* **Remove one of the correlated variables.**
* **Combine the correlated variables into a single variable.**
* **Use regularization techniques like ridge or lasso regression.**
* **Perform principal component analysis to create uncorrelated linear combinations the variables.**

1. **Outlier are data point that deviate from the typical pattern in a dataset. They can have a substantial impact on regression models, affection the regression coefficients, residuals, and the model predictive accuracy.**

**To address outliers, we can**

* **Identify outliers using visualization tools like box plots or scatter plots.**
* **Use statistical test such as the z score or IQR method to detect outliers.**
* **Transform the data such as taking the logarithm of variables to mitigate the impact of outliers.**

1. **Overfitting occurs when a regression model when captures noise or random fluctuation in the training data. Overfit model perform poorly on unseen data set.**
2. **When dealing with non linear relationships, regression models can be enhanced by using polynomial regression.**

**The relationship between predictors and the target variable is modelled as a polynomial function rather then a linear one.**

**To use polynomial regression**

* **You transform the predictors by raising them to different powers to capture the non linear patterns.**