**Machine Learning Assignment**

|  |  |  |  |
| --- | --- | --- | --- |
| (a) | What are the different Machine Learning Paradigms? | | |
| (b) | Differentiate between overfitting and underfitting | | |
| (c) | What is Multivariate Linear Regression? | | |
| (d) | Discuss pair plots with it’s limitations. | | |
| (e) | Differentiate between 2D scatter plot and 3D scatter plot. | | |
| (f) | What is K fold cross validation? | | |
| (g) | What are the different data normalization techniques? | | |
| (h) | What is the difference between balanced and imbalanced data? Discuss with example | | |
| (i) | Differentiate between independent and mutually exclusive events. | | |
| (j) | What is the role of Bayes' theorem in the Naive Bayes classifier? | | |
|  | | |
| 2(a) | | What is logistic regression and how is it different from linear regression? What kind of problems can be solved by logistic regression? |
| (b) | | A company wants to predict sales(in units) based on advertising expenditure(in $).The data for 5 days is   |  |  | | --- | --- | | Advertising($) | Sales(units) | | 100 | 10 | | 200 | 15 | | 300 | 20 | | 400 | 25 | | 500 | 30 |   1.Fit a simple Linear regression model  2.Predict sales if the advertising expenditure is $350  3.Interpret the slope and intercept |

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 3(a) | | A binary classification model is evaluated on a dataset of 100 samples. The confusion matrix is   |  |  |  | | --- | --- | --- | |  | Predicted Positive | Predicted Negative | | Actual Positive | 40 | 10 | | Actual Negative | 5 | 45 |   Calculate accuracy, precision, recall and F1-score.Interprete these metrics in the context of the model | |
| (b) | | Explain the concept of regularization in Linear Regression. Discuss the different techniques in regularization of Linear Regression. | |
|  | |
| 4(a) | What is exploratory data analysis? Compare univariate, bivariate and multivariate analysis with examples. |
| (b) | A dataset has the following points.   |  |  |  | | --- | --- | --- | | X1 | X2 | Class | | 1 | 2 | A | | 2 | 3 | A | | 3 | 3 | B | | 6 | 5 | B |   A new point(4,4) needs to be classified using k=3.Classify the point using Euclidean distance. Re-classify using Manhattan distance and compare the result. |

|  |  |  |  |
| --- | --- | --- | --- |
| 5(a) | | | What is the geometric intuition behind KNN algorithm? Elaborate with an example |
| (b) | | | Construct a box plot for the given data.  4, 8, 6, 5, 3, 2, 7, 9, 10, 12  Identify the median, first quartile (Q1), third quartile (Q3), and the interquartile range (IQR). Determine any outliers. |
|  | | | |
| 6(a) | What is feature creation? Discuss the impact of scale and column standardization. | | |
| (b) | Discuss the various methods of Dimensionality Reduction. | | |

|  |  |  |  |
| --- | --- | --- | --- |
| 7(a) | | | Explain the feature selection using wrapper method with example. |
| (b) | | | Explain binarization and discretization with suitable examples. |
|  | | |
| 8(a) | Consider a dataset with the following attributes for a classification problem where the target variable is whether a person passed (Yes/No) based on their study hours.   |  |  | | --- | --- | | Study Hours | Pass(Yes/No) | | 1 | No | | 2 | No | | 3 | Yes | | 4 | Yes | | 5 | Yes |   Calculate the **entropy** of the target variable. Calculate the **information gain** if you split the data based on the "Study Hours" feature. Based on the information gain, which feature would be selected as the root node for the decision tree? | |
| (b) | Write short notes on 1)Laplace Smoothing 2)Bias Variance trade-off | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 9(a) | A binary classification dataset where the task is to classify whether a person will buy a product (Yes/No) based on their age and income.   |  |  |  | | --- | --- | --- | | Age Group | Income Group | Will Buy Computer(Yes/No) | | Young | High | No | | Young | Low | Yes | | Adult | Low | Yes | | Adult | High | No | | Senior | Low | No | | Senior | High | Yes |   1. Calculate the **prior probabilities** for "Yes" and "No" (Will Buy Product).  2. Calculate the **likelihood** of each feature (Age Group, Income Group) given the classes (Yes/No).  3. Given a new individual with "Adult" age group and "High" income group, calculate the **posterior probability** of the individual buying the product (Yes/No). |
| (b) | Write short notes on1) Information Gain 2) Gini Impurity |