ML ASSIGNMENT 1

1. Define Artificial Intelligence (AI).

Artificial Intelligence is a system that shows behavior that could be interpreted as human Intelligence.

2. Explain the differences between Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), and Data Science (DS).

Artificial Intelligence:

- - Artificial intelligence is a system that shows behaviour that could be interpreted as human intelligence.
- --it can create intelligent machines.
- -- it is a machine that can perform multiple tasks.

Machine Learning:

- -Machine learning is a subset of Artificial intelligence that helps to build the AI applications.
- --It can involves the teaching machines to learn from data so they can make decisions and predictions.

Deep Learning:

- - Deep Learning is a subset of Machine Learning that uses vast volumes of data and complex Algorithms to train a model
- . --deep learning is a advanced form of ml using neural networks with multiple layers.

Data Science :

- - Data Science is a multidisciplinary field that uses scientific methods, processes, algorithms, and systems to extract Knowledge and insights from data.
- --DS uses the ML and DL as tools among other techniques to analyze data.

3. How does AI differ from traditional software development?

AI development differs from traditional software development in that AI systems learn from data and improve over time ,while traditional software fellows explicitly programmed instructions . AI involves training models on large datasets, where as traditional software relies on predefined rules. Ai can adapt to new information dynamically, but traditional software requires manual updates for changes.

4. Provide examples of AI, ML, DL, and DS applications.

- AI Applications :-

1.Voice Assistants :- systems like siri, alexa, and google assistant can understand and respond to spoken commands. 2.autonomous vehicles :- self driving cars 3.Recommendation Systems :- like Netflix, and Spotify use ai. 4.Chatbots

- Machine Learning Applications :

- 1.spam Filtering 2.Fraud Detection 3.Product Recommendations 4.Predictive Maintenance

- Deep Learning Applications:

- 1.Image Recognition 2.Speech Recognition 3.Medical Diagnosis 4.Natural Language Processing

- Data Science Applications:

- 1.Customer Analytics 2.Financial Forcasting 3.Healthcare Analysis 4.Education 5.Sports Analytics

5.Discuss the importance of AI, ML, DL, and DS in today's world.

AI, ML, DL, and DS are pivotal in today's world because they drive innovation and efficiency across industries. AI enhances human capabilities by automating complex tasks, while ML and DL enable systems to learn and adapt, improving decision-making and predictive accuracy. Data Science empowers organizations to extract actionable insights from vast amounts of data, leading to informed strategic decisions and better outcomes. Collectively, these technologies are transforming healthcare, finance, transportation, and countless other fields, fostering advancements that improve quality of life and drive economic growth.

6.What is supervised Learning?

Supervised Learning is a type of machine learning where the model is trained on a labelled dataset, meaning each input data point is paired with the correct output.

7.Provide examples of Supervised Learning Algorithms.

1. Linear Regression. 2. Logistic Regression 3. Decision Trees 4. Support Vector Machines 5. K- Nearest Neighbors (KNN) 6. Naïve Bayes classifier 7. Neural Networks.

8. Explain the process of Supervised Learning.

- Input Data: You start with a set of data where each example is paired with the correct answer. For example, pictures of fruits labeled with their names.
- Training: The computer learns patterns from this labeled data by adjusting its internal settings until it can accurately predict the correct answers for new, unseen data.
- Testing: Once trained, the model is tested on new data to see how well it can predict the correct answers. If it performs well, it's ready to be used for making predictions on similar, unseen data.

9.what are the characteristics of Unsupervised Learning?

Unsupervised Learning in Artificial Intelligence is a type of machine learning that learns from data without human supervision. Unlike supervised learning, unsupervised learning is a machine learning model that is given unlabeled data and allowed to discover patterns and insights without any explicit guidance or instructions.

10. Give examples of Unsupervised Learning algorithms?

K-means :-k-means for clustering problems.

Apriori Algorithm:- Apriori Algorithm for association rule learning problems.

11. Describe Semi- Supervised Learning and its significance?

Semi-Supervised learning is a broad category of machine learning that uses labelled data to ground predictions, and unlabeled data to learn the shape of the larger data distribution.

12. Explain Reinforcement Learning and its applications?

Reinforcement Learning is a sub-field of machine learning which itself is a subfield of artificial intelligence. Applications are Robotics for real time control, Healthcare, Games, Transportation, Energy, Business management, and Financial.

13. How does Reinforcement Learning differ from Supervised and Unsupervised Learning?

Supervised Learning deals with two main tasks: Regression and Classification. Unsupervised Learning deals with clustering and associative rule mining problems. Whereas Reinforcement Learning deals with exploitation or exploration, Markov's decision processes, Policy Learning, Deep Learning and value learning.

14. What is the purpose of the Train-Test-Validation split in Machine Learning?

The purpose of the Train-Test-Validation split in machine learning is to assess the performance and generalisation ability of a model. The training set is used to train the model, the validation set is used to finetune model parameters, and the test is used to evaluate the final performances on unseen data. This separation helps prevent overfitting and provides a reliable estimate of how well the model will perform on new, unseen data.

15. Explain the significance of the training set?

The training set is crucial in machine learning as it serves as the foundation for teaching the model to make predictions. It allows the model to learn patterns and relationships within the data, enabling it to generalise and make accurate predictions

on unseen examples. The quality and diversity of the training set directly impact the model's performance,emphasising the significance of a representative and well-labelled training dataset for successful machine learning outcomes.

16. How do you determine the size of the training, testing, and validation sets?

Determining the size of training , testing, and validation sets involves balancing various factors such as dataset size. Complexity of the problems and computational resources. A common practice is to allocate 70- 80 % of the data to the training set, 10-15% to the training set, and the remaining portion to the validation set.However, This validation can vary depending on the specific requirements of the problem and the available data. Cross-validation techniques can also be employed to ensure robust evaluation of model performance with limited data.

17. what are the consequences of improper Train-Test-Validation split?

Improper Train-Test-Validation split can lead to inaccurate evaluation of a machine learning model's performance and generalisation ability. If the training set is too small, the model may not learn enough patterns, resulting in poor performance. Conversely, if the testing set is too small or not representative, the evaluation may not accurately reflect the model's true performance on unseen data. Additionally, a lock of a validation set can hinder proper tuning of model hyperparameters, leading to suboptimal results.

18. Discuss the trade- offs in selecting appropriate split ratios?

Selecting appropriate split ratios involves trade-offs between having enough data for training, testing, and validation while ensuring a representative sample for evaluation. A larger training set improves model learning but reduces the data available for evaluation, while a smaller training set may lead to overfitting. Finding the right balance depends on the specific characteristics of the dataset and the desired performance of the model.

19. Define model performance in machine learning?

Model performance in machine learning refers to how well a trained model can make predictions or classifications on new, unseen data. It is typically assessed using various metrics such as accuracy, Precision, recall, and F1 score, depending on the nature of the problems. The goal is to maximise the performance metrics to ensure the effectiveness in solving the task it was designed for.

20. How do you measure the performance of a machine learning model?

The performance of a machine learning model is measured using evaluation metrics such as accuracy, Precision, recall, F1 score, and area under the Roc curve (AUC-Roc). These metrics assess the model's ability to make correct predictions or classifications on unseen data and provide insights into its effectiveness in solving the given task.

21. What is overfitting and why is it problematic?

Overfitting occurs when the model cannot generalise and fits too closely to the training dataset instead. Overfitting happens due to several reasons, such as: The training data size is too small and does not contain enough data samples to accurately represent all possible input data values.

22. Provide techniques to address overfitting.

Techniques to address overfitting include:

- a. Regularization : Adding penalty terms to the model's objective function to discourage complex or overly flexible models.
- b. Cross-validation : Splitting the data into multiple subsets for training and validation to assess the model's performance on different data partitions and prevent overfitting.

23. Explain underfitting and its implications.

When a model has not learned the patterns in the training data well and When a model has not learned the patterns in the training data well and is unable to generalise well on the new data, is known as underfitting. An underfitting model has poor performance on the training data and will result in unreliable predictions. Underfitting occurs due to high bias and low variance.

24. How can you prevent underfitting in machine learning models?

There are four ways to prevent the underfitting in the machine learning model --increase the number of features in the dataset. --Increase model complexity. --Reduce noise in the data. -Increase the duration of training the data.

25. Discuss the balance between bias and variance in model performance.

It ensures that we capture the essential patterns inmour model while ignoring the noise present in it. This is called Bias-Variance Tradeoff. It helps optimise the error in our model and keeps it as low as possible.

26. What are the common techniques to handle missing data?

There are two common techniques to handle missing data include:

- a.Imputation: Replacing missing values with a calculated estimate, such as the mean, median, or mode of the available data.
- b.Deletion : Removing observations with missing values, either listwise deletionor pairwise deletion.

27. Explain the Implications of ignoring missing data?

Ignoring missing data can lead to biassed results and reduced model accuracy. It may distort statistics analysis, decrease the representativeness of the sample,and overlook valuable information. Additionally, it can result in inefficient use of available data and potentially undermine the reliability and validity of study findings.

28. Discuss the pros and cons of imputation methods.

1. Pros:

a. Presentation of Data: Imputation allows for the retention of valuable data points, preventing loss of information.

b. Maintain Sample Size: Imputation helps maintain the sample size, which is important for statistics power and generalizability.

2. Cons:

a. Introduction of Bias: Imputation can introduce bias if the imputed values do not accurately represent the true missing values.

b. Assumption Dependency : Imputation methods rely on assumptions about the nature of missingness. Which may not always hold true, leading to erroneous results.

29. How does missing data affect model performance? It can lead to the loss of potentially valuable information from the dataset, Missing values can often be indicative of important patterns in the dataset and ignoring this information can negatively impact the performance of the model. It can reduce the predictive accuracy of the model.

30. Define imbalanced data in the context of machine learning.

A classification data set with skewed class proportions is called imbalanced.Classes that make up a large proportion of the data set are called majority classes. Those that make up a smaller proportion are minority classes.

31. Discuss the challenges posed by imbalanced data.

Imbalanced data poses challenges in machine learning by skewing the model's predictive performance towards the majority class. Leading to poor classification of minority classes. This imbalance can hinder the model's ability to learn and generalise patterns accurately, resulting in biassed predictions and decreased performance on minority class instances.

32. What techniques can be used to address imbalanced data? Imbalanced datasets are used where one class greatly outnumbers others, posing machine learning challenges. To address the techniques like oversampling. Undersampling, SMOTE, ENN, CNN, near miss , and one-sided selection can be employed.

33. Explain the process of up-sampling and down-sampling.

Up sampling is the increasing of the spatial resolution while keeping the 2D representation of an image. Down sampling is the reduction in spatial resolution while keeping the same 2D representation. It is typically used to reduce the storage and/or transmission requirements of images.

34. When would you use up-sampling versus down-sampling?

Up-sampling used for may be more effective in improving model performance.

--use up-sampling when you want to increase the representation of the minority class. This can be helpful for models that struggle to learn from rare events.

--If the goal is to improve the model efficiency or reduce the risk of overfitting, down-sampling may be a better option.

35. What is SMOTE and how does it work?

SMOTE stands for Synthetic Minority Oversampling Technique. SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It works by interpolating between existing minority class data points, essentially creating new data points along the lines connecting them in feature space.

36. Explain the role of SMOTE in handling imbalanced data

SMOTE stands for Synthetic Minority Oversampling Technique. It is one of the most commonly used oversampling methods to solve the imbalance problem.It aims to

balance class distribution by randomly increasing minority class examples by replicating them.

37. Discuss the advantages and limitations of SMOTE.

- Advantages: - Boosts minority Classes - Effective handling of imbalanced data - Preservation of information - Reduction of overfitting
- Limitations: - Generation of Synthetic Data - Sensitivity to Noisy data. - Performance in High Dimensional Spaces - Impact on computational Resources

38. Provide examples of scenarios where SMOTE is beneficial.

-- Credit card Fraud Detection -- Medical Diagnosis -- Customer Churn Prediction -- Anomaly Detection

39. Define data interpolation and its purpose.

Data Interpolation:-data interpolation is the process of using known data values to estimate unknown data values is called data interpolation. It is mostly used to predict the unknown values for any geographical related data points such as noise level, rainfall, elevation, and so on.

40. What are the common methods of data interpolation?

Common methods of data interpolation: a. Linear interpolation b. Polynomial interpolation c. Spline interpolation d. Nearest Neighbor interpolation e. Inverse Distance Weighting.

41. Discuss the implications of using data interpolation in machine learning.

--injects assumptions --Reduced accuracy -- impact on model performance -- Data quality and robustness -- Bias and overfitting -- complexity and computational Resources

42. What are outliers in a dataset?An outlier is a data point in a data set that is distant from all other observations. A data point that lies outside the overall distribution of the dataset.

43. Explain the impact of outliers on machine learning models.

It causes various problems during our statistical analysis. It may cause a significant impact on the mean and the standard deviation.

44. Discuss techniques for identifying outliers.

- --Visualisation Technique :-a) Box plot, b) Scatter plot, c)Histogram
- -- Statistical Methods :- a) Z-Score, b) Modified Z-Score, c) Percentile Method
- -- Machine Learning Technique :- a) Isolation Forest, b) Local Outlier Factor(LOF)

45. How can outliers be handled in a dataset?

1) reducing the weights of outliers (trimming weight)
2) changing the values of outliers (Winsorization, trimming, imputation)
3) using robust estimation techniques (M-estimation).

46. Compare and contrast Filter, Wrapper, and embedded methods for feature selection.

1. Filter Methods:

Pros: Quick, simple, and computationally efficient.
Cons: Ignore interactions between features.
Example: Chi-square test, ANOVA.

2. Wrapper Methods:

Pros: Consider feature interactions and yield high accuracy.
Cons: Computationally intensive and risk overfitting.
Example: Recursive Feature Elimination (RFE), Forward Selection.

3. Embedded Methods:

Pros: Balance between filter and wrapper methods by incorporating feature selection during model training.
Cons: Dependent on the chosen model and may be complex.
Example: Lasso (L1 regularisation), Decision Tree feature importance.

47. Provide examples of algorithms associated with each other.

- Filter Methods: Examples include Chi-square test, ANOVA, and mutual information.
- Wrapper Methods: Examples include Recursive Feature Elimination (RFE) and Forward/Backward Feature Selection.
- Embedded Methods: Examples include Lasso (L1 regularization), Ridge (L2 regularization), and Decision Trees with built-in feature importance.

48. Discuss the advantages and disadvantages of each feature selection method.

1. Filter Methods:

Advantages: Fast and computationally efficient.
Disadvantages: Ignores feature interactions, which can reduce accuracy.

2. Wrapper Methods:

Advantages: Consider feature interactions, leading to higher accuracy.
Disadvantages: Computationally intensive and prone to overfitting.

3. Embedded Methods:

Advantages: Integrate feature selection into model training, balancing accuracy and efficiency. Disadvantages: Model-dependent and can be complex to implement.

49. Explain the concept of feature scaling. Feature scaling is the process of normalising or standardising the range of independent variables or features in a dataset to ensure they contribute equally to the model's performance, enhancing convergence and accuracy.

50. Describe the process of feature scaling.

Feature scaling involves transforming features to a common scale, typically using methods like Min-Max normalisation (rescaling values to a range of [0, 1]) or Standardization (scaling to zero mean and unit variance).

51. How does mean normalization differ from standardisation?

Mean normalisation scales features to a range centred around zero using the formula (x−mean)/(max−min) while standardisation transforms features to have a mean of zero and a standard deviation of one using (x−mean)/std.

52. Discuss the advantages and disadvantages of min-max scaling.

Advantages: Min-max scaling preserves relationships in the data and bounds features within a fixed range, which is useful for algorithms sensitive to feature magnitude.

Disadvantages: It is sensitive to outliers, which can distort the scaling and affect model performance.

53. What is the purpose of unit vector scaling? The purpose of unit vector scaling is to transform feature vectors to have a unit norm (length of 1), ensuring that the magnitude of the vector is normalised, which helps in algorithms where direction is more important than magnitude, such as in cosine similarity.

54. Define Principle Component Analysis(PCA).

Principal Component Analysis (PCA) reduces dimensionality by transforming data into orthogonal components ordered by explained variance, retaining essential information.

55. Explain the steps involved in pca.

PCA Steps:

1. Standardization: Normalize features to have zero mean and unit variance.

2. Compute Covariance Matrix: Calculate the covariance matrix of the standardised data.
3. Eigendecomposition: Find eigenvectors and eigenvalues of the covariance matrix.
4. Select Principal Components: Choose top eigenvectors based on explained variance.
5. Project Data: Transform original data onto the new feature space defined by principal components.

56. Discuss significance of eigenvalues and eigenvectors in pca.

Eigenvalues represent the amount of variance explained by each principal component, eigenvectors indicate the direction of maximum variance in the dataset, crucial for reducing dimensions while retaining essential information in PCA.

57. How does PCA help in dimensionality reduction?

PCA reduces the dimensionality of a dataset by transforming it into a new set of orthogonal features (principal components) that capture the maximum variance, allowing for the retention of most relevant information while discarding less important features.

58. Define data encoding and its importance in machine learning.

Data encoding is the process of converting categorical data into a numerical format, crucial for machine learning algorithms to interpret and analyse non-numeric features effectively, enhancing model performance and accuracy.

59. Explain Nominal Encoding and provide an example.

Nominal Encoding converts categorical variables into binary features, assigning a unique binary code to each category. For instance, in the "color" feature with categories "red," "blue," and "green," nominal encoding would represent them as [1, 0, 0], [0, 1, 0], and [0, 0, 1], respectively.

60. Discuss the process of One Hot Encoding.

One Hot Encoding converts categorical variables into binary vectors where each category is represented by a binary digit, with only one bit set to 1 indicating the presence of that category while others are set to 0.

61. How do you handle multiple categories in one hot encoding?

In One Hot Encoding, each category is represented by a binary vector, with a separate binary feature for each category. When there are multiple categories, each category is assigned a separate binary vector, with only one bit set to 1 to indicate the presence of that category. This approach ensures that each category is uniquely represented and prevents any ordinal relationship among categories.

62. Explain mean encoding and its advantages.

Mean encoding, also known as target encoding, replaces categorical values with the mean of the target variable for each category.
advantages include capturing target-specific information, reducing dimensionality, and maintaining the ordinality of categories, making it useful for tree-based models and linear models with high-cardinality categorical features. However, it is prone to overfitting, especially with small datasets or rare categories.

63. Provide examples of ordinal Encoding and label encoding.

Ordinal Encoding assigns a unique integer to each category based on their order, such as "low" as 0, "medium" as 1, and "high" as 2.Label Encoding converts categorical labels into integer values, like "red" as 0, "blue" as 1, and "green" as 2.

64. What is Target Guided Ordinal Encoding and how is it used?

Target Guided Ordinal Encoding assigns ranks to categories based on the mean of the target variable, providing ordinality while considering the target's influence, commonly used in classification tasks to encode categorical variables.

65. Define covariance and its significance in statistics.

Covariance measures the degree to which two variables change together, indicating the direction of linear relationship between them; it is crucial in statistics for understanding the relationship and variability between variables.

66. Explain the process of correlation check.

Correlation check involves calculating correlation coefficients (such as Pearson, Spearman, or Kendall) between pairs of variables to assess the strength and direction of their linear relationship, helping to identify patterns and dependencies in the data.

67.what is the pearson correlation coefficient?

The Pearson correlation coefficient measures the linear relationship between two continuous variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear correlation.

68. How does Spearman’s Rank correlation differ from Pearson’s Correlation?

Spearman's Rank correlation assesses the strength and direction of monotonic relationships between variables by ranking the data, making it robust to outliers and suitable for ordinal or non-normally distributed data, whereas Pearson's correlation evaluates linear relationships between continuous variables using raw data values.

69. Discuss the importance of variance inflation Factor (VIF) in feature selection.

The Variance Inflation Factor (VIF) is important in feature selection as it helps identify multicollinearity among predictor variables, guiding the removal of highly correlated features to improve model interpretability and stability.

70. Define feature selection and its purpose.

Feature selection is the process of identifying and selecting a subset of relevant features from a dataset to improve model performance, reduce overfitting, enhance interpretability, and decrease computational complexity.

71.Explain the process of Recursive Feature Elimination.

Recursive Feature Elimination (RFE) recursively removes less important features based on model performance until the optimal subset is achieved, effectively ranking features by their contribution to model accuracy.

72.How does Backward Elimination work?

Backward Elimination starts with all features included in the model, iteratively removing the least significant feature based on a chosen criterion (e.g., p-value) until no further improvement is observed in model performance, aiming to find the most parsimonious model.

73. Discuss the advantages and limitations of Forward Elimination.

Forward Elimination systematically adds features to the model, simplifying tForward Elimination systematically adds features to the model, simplifying the selection process and reducing computational complexity, but it may overlook interactions between features and result in suboptimal models if feature importance changes during the selection process.

74.What is feature engineering and why is it important.

Feature engineering involves creating new features or transforming existing ones to improve model performance and interpretability, crucial for optimising model accuracy and extracting meaningful insights from data.

75.Discuss the steps involved in feature engineering.

Feature engineering steps include:

1. Feature Creation: Generating new features from existing ones.

2. Feature Transformation: Scaling, normalizing, or encoding features to improve model performance.
3. Feature Selection: Identifying and selecting relevant features to reduce dimensionality and improve model interpretability.

76.Provide examples of feature engineering techniques.

Polynomial Features: Creating new features by raising existing ones to higher powers. Interaction Features: Multiplying or combining two or more features to capture their joint effect on the target variable.

77. How does feature selection differ from feature engineering?

Feature selection involves choosing a subset of existing features, while feature engineering involves creating new features or transforming existing ones, both aimed at improving model performance and interpretability.

78. Explain the importance of feature selection in machine learning pipelines.

Feature selection is crucial in machine learning pipelines as it reduces overfitting, improves model interpretability, enhances computational efficiency, and facilitates better generalisation to unseen data by focusing on the most relevant features.

79. Discuss the impact of feature selection on model performances.

Feature selection can significantly improve model performance by reducing overfitting, decreasing computational complexity, enhancing interpretability, and increasing the model's generalisation ability to new data.

80. How do you determine which features to include in a machine learning model.

Features are typically selected based on their relevance to the target variable, using techniques like statistical tests, feature importance rankings from models, domain knowledge, or automatic selection algorithms to identify the most informative ones for the model.