ML Assignment - 1

Q1. Define Artificial intelligence AI?

Ans:

Smart application that can perform own task without any human intervention.

Ex: Robot, self-driving car.

Q2. Explain the differences between Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), and Data Science (DS).

Ans:

Artificial Intelligence (AI):

smart application that can perform own task without human intervention.

Ex: Robot, self-Driving car

Machine Learning:

Machine learn pattern from data and tried to replicate same in future

Ex: Spam & Ham, Diabetes or not, price of house

Deep Learning:

Specialized machine learning algorithms that mimic human brain

Ex: chat gpt, Devin AI, Object detections, Image recommendation

Data Science:

Interdisciplinary field comprised of SC, math and domain knowledge.

Q3. How does AI differ from traditional software developments?

Ans:

Traditional software is programmed to perform tasks.AI is program to learn perform the task.

code is primarily artifact traditional software. IN AI software Code is not primarily artifact. when we build Ai Software we have to write code.

Q4. Provide examples of AI, ML, DL, and DS applications?

Ans:

AI is used in virtual assistants, recommendation systems, and more.

ML is applied in image recognition, spam filtering, and other data tasks.

ML Assignment - 1

DL is utilized in autonomous vehicles, speech recognition, and advanced AI applications.

DS is utilized In Search Engines, Transport, Finance, Ecommerce.

Q5. Discuss the importance of AI, ML, DL, and DS in today's World?

Ans:

Artificial Intelligence (AI) sets the stage for machines that can simulate human intelligence. Machine Learning (ML) evolves from AI, giving machines the ability to learn and grow from experience. Deep Learning (DL), nestled within ML, drives machines to understand and operate on a level akin to human intuition.

Q6. what is supervised Learning?

Ans:

In supervised Learning We have two problem statements:

1. Regression
2. classification

In supervised machine learning we have given target variable, independent variable, labelled, tag.

supervised machine learning input variable is explained relationalship between output variable.

Q7. Provide the examples of supervised learning algorithms?

Ans:

The most commonly used Supervised Learning algorithms are decision tree, logistic regression, linear regression, support vector machine.

Example: supervised learning problems is predicting house prices, diabetic or not, spam or ham.

Q8. Explain the process of supervised Learning?

Ans:

Supervised machine learning algorithm that used labelled dataset which has dependant variable.

it is used to explain the relationship between input and output variable.

ML Assignment - 1

Q9. what are the characteristics of unsupervised Learning?

Ans:

In unsupervised ML No historical data, Not dependant variable.

In unsupervised ML data contain similar groups

Ex:

Mall ==> Shirt

Jacket

Jeans

Q10. Give examples of unsupervised Learning algorithms?

Ans:

1. clustering
2. Fraud detection
3. dimensionality reduction
4. anomaly detection

Q11. Describe semi-supervised Learning and its significance?

Ans:

In semi supervised learning it is combination of supervised and unsupervised learning.

in semi supervised learning used labelled or unlabelled data. In target variable we have continuous

or discrete both data is available.

this can be especially helpful in domains where expert labelling is difficult, such as medical diagnosis or anomaly detection.

Q12. Explain Reinforcement Learning and its applications?

Ans:

Reinforcement learning is an area of ml concerned with how intelligent agent ought to take action in an environment.

Reinforcement learning (RL) is a machine learning (ML) technique that trains software to make decisions to achieve the most optimal. results. It mimics the trial-and-error learning process that humans use to achieve their goals.

ML Assignment - 1

Q13. How does Reinforcement Learning differ from supervised and unsupervised Learning?

Ans:

Supervised learning:

supervised leaning used labelled data to learn pattern or output of data.

Unsupervised Learning:

unsupervised learning uses unlabelled data to learn pattern from data.

Reinforcement learning:

Reinforcement learning (RL) is a machine learning (ML) technique that trains software to make decisions to achieve the most optimal results. It mimics the trial-and-error learning process that humans use to achieve their goals.

Q14 what is the purpose of the Train-Test-Validation split in Machine learning ?

Ans:

*** Purpose of train test & validation

Train:

train data used for training of model.

Test:

Test data is used for testing the model accuracy

Validation:

Validation data is used for hyperparameter tuning.

Q15. Explain the significance of the training set?

Ans:

Training set is subset of data which is used to train machine learning model learn pattern from data. Training data is used to learn the pattern relational ship & nuance from the data. training set is largest subset of data is used to train the model.

ML Assignment - 1

Q16.How do you determine the size of the training, testing, and validation sets?

Ans:

Train:

train data size is 60 %. we can train the model using 60% data.

Test:

Test data size is 20%. We can test the model accuracy using these specific data.

Validation data:

Validation data size is 20%. we have to use the validation data for hyperparameter tuning.

Q17 what are the consequences of improper Train-Test-Validation splits?

Ans:

Incorrectly shuffling or sorting the data before splitting can introduce bias and affect the generalization of the final model. For example, if the dataset is not shuffled randomly before splitting into training set and validation set, it may introduce biases or patterns that the model can exploit during training.

Q18 Discuss the trade-offs in selecting appropriate split Ratios?

Ans:

In statistics and machine learning, the bias–variance trade-off describes the relationship between a model's complexity, the accuracy of its predictions, and how well it can make predictions on previously unseen data that were not used to train the model.

Q19 Define model performance in machine learning?

Ans:

Model performance in machine learning is how the model giving accurate prediction on unseen data that term is called as model performance typically measure model performance using a test set, where you compare the predictions on the test set to the actual outcomes.

Q20. How do you measure the performance of a machine learning model?

Ans:

The machine learning model can be evaluated using metrics that analyse its ability to make predictions, its weaknesses, and how well it can generalize future predictions.

ML Assignment - 1

For Regression:

we have use

1. mean square error
2. mean absolute error
3. Root mean square error
4. r2 score

for classification:

1.confusion matrix

2.accuracy score

3.classification report

Q21 what is overfitting and why is it problematic?

Ans:

Model performing well during training data but worst perform during testing that is called as overfitting

Train ==> while training accuracy is high

test ===> while testing accuracy is low

Problem due to overfitting:

In ml while training accuracy is high & testing accuracy is low that is overfitting. overfitting can reduce the model reliability; it can produce inaccurate prediction.

Q22 Provide techniques to address overfitting?

Ans:

1.Cross validation

2.Evaluation

3.Training early stopping

ML Assignment - 1

Q23 Explain underfitting and its implications?

Ans:

Basically, Low Train accuracy & low-test accuracy is underfitting. Underfitting happen because of

where data model is unable to capture the relationship of input and output variable. it causes high error while training.

Q24 How can you prevent underfitting in machine learning models?

Ans:

Increase model complexity.

Increase the number of features, performing feature engineering.

Remove noise from the data.

Increase the number of epochs or increase the duration of training to get better.

Q25 Discuss the balance between bias and variance in model performance?

Ans:

High Bias & Low Bias:

1. Low training error is also known as low bias
2. high training error is known as high bias

High Variance & Low Variance:

1. low testing error is known as low variance
2. high testing error is known as high variance

In case:

Overfitting:

1. training accuracy is high ==> low bias
2. testing accuracy is low ==> high variance

ML Assignment - 1

Underfitting:

1. training accuracy is low ==> high bias
2. testing accuracy is low ==> high variance

Generalize model:

1. training accuracy is high ==> low bias
2. testing accuracy is high ==> low variance

Q26 what are the common techniques to handle missing data?

Ans:

1. Missing Completely at random
2. Missing At Random
3. Missing not at Random

We can handle missing data using Imputation.

Numerical Data: we will use mean & median imputation

Categorical Data: we will use mode imputation.

Q27. Explain the implications of ignoring missing data?

Ans:

Implication of Missing Data:

1. sampling Bias
2. Missing data implicate on model performance
3. not getting generalize model
4. problematic to find valuable insight from data.

ML Assignment - 1

Q28. Discuss the pros and cons of imputation methods?

Ans:

Pros & cons of imputation method

1. Numerical feature: if outlier treatment is done, we can use mean imputation

   if outlier treatment has not done, we can use median imputation,

2. categorical feature:

   here we can use mode imputation.

Q29 How does missing data affect model Performance?

Ans:

Due to missing data model will predict incorrect results. Or sometime model will not understand data.

Q30. Define imbalanced data in the context of machine learning?

Ans:

When one class has higher percentage of data as compare to another class is called as imbalance data

example:

Majority class ==> 90

Minority class ==> 10

Q31 Discuss the challenges posed by imbalanced data?

Ans:

Imbalance data can cause problem like biased model, inaccurate prediction. In case of classification

when one class is majority class & another class is Minority class so the classification result will be majority.

ML Assignment - 1

Q32 What techniques can be used to address imbalanced data?

Ans:

There are three techniques to address imbalance data:

1.oversampling

2.undersampling

3.SMOTE (synthetic minority oversampling techniques.

Q33 Explain the process of up-sampling and down -sampling?

Ans:

Up sampling:

In up sampling repeat the data from majority class which is equivalent to minority class

down Sampling:

In down sampling use the data from majority class which equivalent to minority class.

Q34 When would you use up-sampling versus down-sampling?

Ans:

If the dataset has an imbalanced distribution, up sampling may be more effective in addressing the problem. Down sampling can result in a loss of important information and may not be effective in balancing the class distribution.

Q35 What is SMOTE and How does it work?

Ans:

In SMOTE technique state that same nature of data will be part of same group. SMOTE is an oversampling technique that generates synthetic samples from the minority class. It obtains a synthetically class-balanced or nearly class-balanced training set, then trains the classifier.

ML Assignment - 1

Q36 Explain the role of SMOTE in handling imbalanced data?

Ans:

SMOTE (synthetic minority oversampling technique) 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. SMOTE synthesises new minority instances between existing minority instances.

Q37. Discuss the advantages and limitations of SMOTE?

Ans:

A drawback of SMOTE is that it doesn't consider the majority class while creating synthetic samples. Additionally, it can create synthetic samples between samples that represent noise. As a result, the augmented dataset will have more noise than the original one, which can hurt performance.

While SMOTE is highly effective, it's not without its challenges and limitations: Data Quality: SMOTE assumes that the minority class instances are close in feature space. If the minority class is very sparse or if the data quality is poor, the synthetic samples created may not be representative.

Q38 Provide examples of scenarios where SMOTE is beneficial?

Ans:

SMOTE creates new synthetic spam emails based on existing ones, balancing the dataset for better spam detection.

Q39 Define data interpolation and its purpose?

Ans:

Interpolation is a process of determining the unknown values that lie in between the known data points. It is mostly used to predict the unknown values for any geographical related data points such as noise level, rainfall, elevation.

Q40 What are the common methods of data interpolations?

Ans:

There are three common method of data interpolation :

1. Linear interpolation
2. cubic interpolation
3. Polynomial interpolation

ML Assignment - 1

Q41 Discuss the implications of using data interpolation in machine learning?

Ans:

interpolation refers to the process of estimating unknown values that fall between known data points. This can be useful in various scenarios, such as filling in missing values in a dataset or generating new data points to smooth out a curve.

Q42. What are outliers in a dataset?

Ans:

Outlier in dataset is most extreme values are present in dataset is called outliers.

Q43 Explain the impact of outliers in machine learning models?

Ans:

Outlier impact in machine learning model:

1.Model can be biased due to outliers

2.model can't give accurate prediction.

3.reliability

Q44 Discuss techniques for identifying outliers?

Ans:

We can find outlier using:

1. Box plot

2. python describe function

Q45 How can outliers be handled in a dataset?

Ans:

1. If we want to removed extreme values from data we need to use 5 point summery for outlier treatment

we use IQR range find outliers & removed it.

2. if we don't want remove outlier from data, we can impute it with mean and median.

ML Assignment - 1

Q46 Compare and contrast Filter, Wrapper, and Embedded methods for feature selection?

Ans:

Filter methods:

Evaluate features independently of the model, based on their individual characteristics and statistical tests. They're fast and good for large datasets, but don't consider feature relationships. Examples include variance thresholds and information gain.

Wrapper methods:

Test different combinations of features to see which performs best for a specific model. They can offer better model performance, but are computationally intensive and time-consuming. Examples include sequential feature selection.

Embedded methods;

Select features while training the model itself. They balance efficiency and model-specific learning, and are less prone to overfitting. Examples include LASSO and Ridge.

Q47 Provide examples of algorithms associated with each method?

Ans:

There are 3 algorithms associated with these method

1.Ridge regression (l1 regularization)

2.Lasso regression (l2 regularization)

3.Elastic net regression (L1 + l2 regularization)

Q48 Discuss the advantages and disadvantages of each feature selection method?

Ans:

Filter Method:

Advantage:

1.computationaly efficient, fast processing, less prone to overfitting.

Disadvantage:

1.less precision could fail to find the feature.

ML Assignment - 1

Wrapper method:

Advantage:

1.high precision

disadvantage:

1.computationally expensive, less precision, tend to overfitting

Embedded method:

Advantage:

1.Low extra cost

Disadvantage:

1.Learning dependant

Q49 Explain the concept of feature scaling?

Ans:

Feature scaling state that the feature will bring on same scale for modelling purpose.

Q50 Describe the process of standardization?

Ans:

Standardization is a statistical technique used in data preprocessing to make different variables more comparable. It's like translating all these different data “languages” into one universal dialect.

Q51 How does mean normalization differ from standardization ?

Ans:

Normalization:

In normalization minimum & maximum value are used for scaling.

Standardization:

In standardization mean & standard deviation is used for scaling.

ML Assignment - 1

Q52 Discuss the advantages and disadvantages of Min-Max scaling.

Ans:

Advantage:

1.Simple to understand and implement.

2.Effective for a wide range of data.

Disadvantage:

1.Sensitive to outlier.

2.can lead to information loss.

Q53 What is the Purpose of unit vector scaling?

Ans:

Unit vector scaling is a technique that can be used to normalize the range of independent variables or features of data. It's also known as feature scaling and is often performed during data preprocessing.

Q54 Define Principal Component Analysis (PCA)?

Ans:

Principal component analysis (PCA) is a dimensionality reduction and machine learning method used to simplify a large data set into a smaller set while still maintaining significant patterns and trends.

Q55 Explain the steps involved in PCA?

Ans:

1.Standardize the range of continuous initial variables.

2.Compute the covariance matrix to identify correlations.

3.Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components.

4.Create a feature vector to decide which principal components to keep.

5.Recast the data along the principal components axes.

ML Assignment - 1

Q56 Discuss the significance of eigenvalues and eigenvectors in PCA.

Ans:

To determine the most significant principal components, you can rank the eigenvectors from highest to lowest eigenvalue. The first eigenvector is the main principal component, and subsequent eigenvectors are orthogonal to it so that they can span the entire x-y area.

Q57 How does PCA helps in dimensionality reduction?

Ans:

Principal component analysis (PCA) is a linear dimensionality reduction technique that reduces the number of dimensions in a data set while retaining most of its information. It does this by transforming the original variables into a smaller set of new, uncorrelated variables called principal components.

Q58 Define data encoding and its importance in machine learning.

Ans:

Data encoding is the process of converting data, usually categorical or text data, into a numerical format that machines can understand and process. It's a crucial step in preparing data for machine learning algorithms, as these algorithms primarily work with numerical data.

Q59 Explain Nominal Encoding and provide an example.

Ans:

In nominal encoding convert categorical data into numerical. no order or rank in the data.

Example: marital status, gender

Q60 Discuss the process of One Hot Encoding?

Ans:

- In onehot Encoding Categorical data convert into numerical data.

- There is no order in data.

Example: Single, Married, In relationship

single married in relationship

1 0 0

0 1 0

0 0 1

Q61 How do you handle multiple categories in One Hot Encoding?

Ans:

One-hot encoding is a technique that converts categorical variables into numerical variables by creating a new binary variable for each possible value of the categorical variable. When dealing with multiple categories, you can:

Limit encoding to frequent labels Encode only the 10 most frequent labels, and group all other labels under a new category. Drop a dummy variable Since dummy variables contain redundant information, you can drop one of the newly created columns to avoid the dummy variable trap.

Q62 Explain Mean Encoding and its advantages?

Ans:

Target encoding, also known as mean encoding, involves replacing each category with the mean (or some other statistic) of the target variable for that category. Here's how target encoding works: Calculate the mean of the target variable for each category. Replace the category with its corresponding mean value.

Q63 Provide examples of Ordinal Encoding and Label Encoding?

Ans:

Assignee numerical label to each category.

Example:

colors: red, green, blue

Q64 What is Target Guided Ordinal Encoding and how is it used?

Ans:

Target-guided ordinal encoding is a technique used to encode categorical variables for machine learning models. This encoding technique is particularly useful when the target variable is ordinal, meaning that it has a natural order, such as low, medium, and high.

ML Assignment - 1

Q65 Define covariance and its significance in statistics.

Ans:

covariance is a statistical tool that measures the relationship between two random variables and how much they change together. It can indicate the direction of a relationship, such as whether the variables tend to move in tandem or show an inverse relationship, but it does not indicate the strength of the relationship or the dependency between the variables.

Q66 Explain the process of correlation check?

Ans:

Correlation can be check between two random variable it should be positive or negative

by using heat map or corr() method in python we can check linear relationship between two feature.

Q67 What is the Pearson Correlation Coefficient?

Ans:

The Pearson correlation coefficient (r) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

Q68 How does Spearman’s Rank Correlation differ from Pearson's Correlation?

Ans:

Spearman's rank correlation

A nonparametric measure that assesses how well a monotonic function describes the relationship between two variables. It's often used for ordinal data, nonnormally distributed continuous data, or data with outliers. Spearman's correlation orders values from highest to lowest, without considering the distances between them. A perfect Spearman correlation of +1 or −1 occurs when each variable is a perfect monotone function of the other.

Pearson's correlation

Assesses linear relationships between quantitative variables that follow a normal distribution. It's typically used for jointly normally distributed data. Pearson's correlation compares the mean value of the product of the standard scores of matched pairs of observations. Positive values indicate a positive correlation, negative values indicate a negative correlation, and zero values indicate no correlation.

ML Assignment - 1

Q69 Discuss the importance of Variance Inflation Factor (VIF) in feature selection?

Ans:

The variance inflation factor (VIF) is a statistical tool that measures how much the variance of an estimated regression coefficient is increased due to collinearity. It's a useful tool for feature selection in machine learning analysis because it can help identify and eliminate variables that cause multicollinearity. Multicollinearity can lead to unstable parameter estimation, weak predictive ability, and less dependable statistical conclusions. Removing multicollinearity can improve the accuracy and sensitivity of classification models, and the stability and generalization performance of extreme learning machines (ELM) model.

Q70 Define feature selection and its purpose?

Ans:

The goal of feature selection is to find the best set of features from the available data that models the given problem to yield a machine learning model with good performance and robustness.

Q71 Explain the process of Recursive Feature Elimination?

Ans:

Recursive Feature Elimination is a feature selection method to identify a dataset's key features. The process involves developing a model with the remaining features after repeatedly removing the least significant parts until the desired number of features is obtained.

Q72 How does Backward Elimination work?

Ans:

Backward elimination is a more systematic approach that starts with a complete set of features and removes features one by one until the model performance reaches a peak. This method is more computationally efficient but may not find the optimal set of features either.

Q73 Discuss the advantages and limitations of Forward Elimination?

Ans:

Advantages:

1.The most fundamental solution algorithm.

2.Basis for computing inverse; can solve multiple sets of equations.

limitation:

1.Solution of one set of linear equations at a time.

2.Less efficient for a single set of equations.

Q74 What is feature engineering and why is it important?

Ans:

Feature engineering is a crucial step in machine learning that involves selecting and transforming raw data into features that can better represent an underlying problem for a predictive model. The goal is to improve model accuracy by providing more relevant and meaningful information.

Q75 Discuss the steps involved in feature engineering?

Ans: Step in feature engineering

1. deal with unique value

2. handling Nan values

3. handling outliers

4. encoding

5. aggregate feature

6. feature selection

7. EDA

Q76 Provide example of feature engineering techniques.

Ans:

1.Feature extraction

2.Imputation

3.scaling

4. handling missing values

5.handling outliers

6.onehot encoding

7.log transform

ML Assignment - 1

Q77 How does feature selection differ from feature engineering?

Ans:

Feature engineering and feature selection are both important techniques used in machine learning to improve model accuracy and performance. They have different objectives, but they can overlap and are often used together in a workflow.

Q78 Explain the importance of feature selection in machine learning pipelines?

Ans:

In the machine learning process, feature selection is used to make the process more accurate. It also increases the prediction power of the algorithms by selecting the most critical variables and eliminating the redundant and irrelevant ones. This is why feature selection is important.

Q79 Discuss the impact of feature selection on model performance?

Ans:

In the machine learning process, feature selection is used to make the process more accurate. It also increases the prediction power of the algorithms by selecting the most critical variables and eliminating the redundant and irrelevant ones. This is why feature selection is important.

Q80 How do you determine which features to include in a machine-learning model?

Ans:

For feature selection process we have used 2 method VIF Method and RFE method to select best feature for model.