Sister Nivedita University

Name: Farhad Dubey

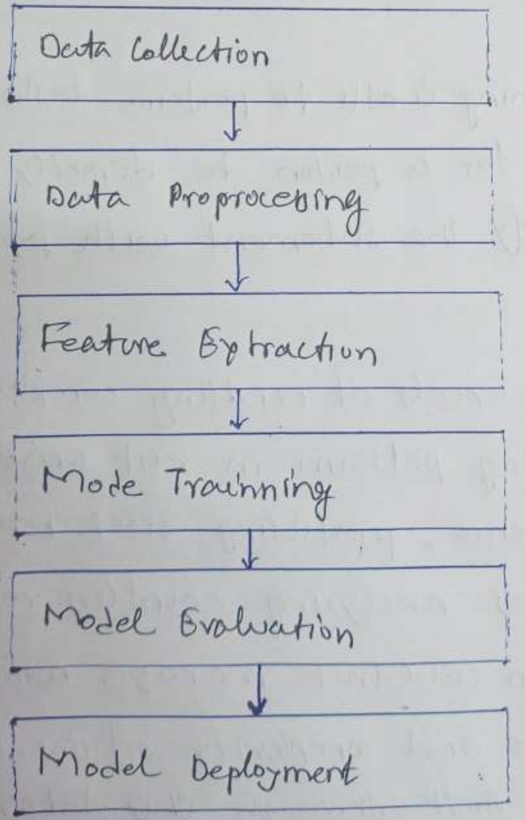
Department: CSE Sec: B

Roll No.: 2111200001217 Subject: Machine Learning.

Reg No.: 210012175539

Q1. Draw a simple ML work with a suitable block diagram. State the features of ML.

Ans: Here's a simple block diagram illustrating a ML workflow:



> Features of Machine Learning :

1. Automated learning: ML algorithms improves automatically through experience.

2. Adeptability: Capable of adopting to new data without manual intervention.

3. Prediction: Can make predictions based on learned patterns.

Q2. 'The Machine Learning is able to perform tasks that are two complex for a person to directly increment! - justify the statement with proper examples.

> Machine Learning excels at handling complex tasks by recognizing patterns in vast amount of data. For instance, predicting stock market movements involves analyzing countless variables, something a person couldn't manage manually due to the volume and complexity of the data. ML algorithms can sift through this data,

identify pattern and make more accurate predictions, demonstrating their capability to handle tasks beyond human capacity.

3. Give one example of dependent, independent & demographic variable.

> Dependent variable: Sales Revenue, House Price

Independent variable: Ads Spend, Square Footage

Demographic variable: Age, Location

4. "Classification is better than Regression" – Justify the statement with proper example:

> In ML, saying "Classification is better than Regression" is not universally true. It depends on the problem that's being tried to solve.

Classification: In a dataset of emails, some of which are spam & some are not.
Here the goal is to classify emails into "spam"

or "not spam". Here the output is categorical, where classification is better than regression.

Value Prediction: Predicting house prices based on features like square footage, no of bedrooms & location. Here a continuous value (house price) is being predicted based on inputs. Here classification is totally not applicable, hence Regression is the only option.

So, the above statement is only valid when the output prediction is categorical.

5.
Ans: Suitable ML model for this database would be Multiple Linear Regression. Because:

1. Predicting a Numerical value: Since we want to predict a numerical value, Regression models are best for it.

2. Multiple Features: We have multiple input features, making it suitable for multiple linear regression.

3. Simple and Interpretable: multiple linear regression provides a clear interpretation of how each feature influences the prediction, which can be beneficial for understanding the underlying relationship in the data.

6. Ans: Both Clustering & Association are unsupervised learning techniques.

Clustering is used to group similar data points together based on certain feature without prior knowledge of labels.

Association is used to discover interesting relations between variables in large datasets.

- Clustering is preferred when:
  - Exploring unknown patterns or segmenting data

- No predefined labels are available.

Ans. 7 The robotic dog has been trained extensively for five days to learn arm movements and complete tasks on time. However, there's no mention of training & navigating around specific individuals.

Hence it will not be able to identify and navigate around a specific person without additional training or program for facial recognition. Hence it will reject the task.

8. Because, better synthesis and generation of complex data became possible with the introduction of advance machine learning algorithms and high-performance computing thats introduced AI

--end--.

9. key applications of ML are:

1. Healthcare: ML assists in disease identification, personalized treatment recommendations, and predicting patient outcomes based on historical data.

2. Finance: ML algorithm predict stock prices, detect fraudulent transactions and asses credit risks by analyzing vast amount of financial data.

3. E-commerce: Recommender systems use ML to personalize product recommendations based on user behaviour and preferences.

4. Natural Language Processing

5. Autonomous vehicle

6. Image & Speech Recognition.

10. ML algorithms detect email spam and malware by analysing features like content, sender information and file behaviour.

↳ Email Spam Filtering:

1. Feature Extraction: Extracts keywords & sender details.

2. Model Training: Trained by algorithms like Naive Bayes, SVM on labeled data.

3. Prediction: Classifying incoming emails as spam or not.

↳ Malware Filtration:

1. File Analysis: Examine file types and behaviour.

2. Model training: trained by algos like Decision tree on labeled malware samples.

3. Detection; Flag suspicious files in emails.

11. 1. Data Collection.

2. Data Preprocessing

3. Feature Engineering.

4. Model Selection.

5. Model training.

6. Model Evaluati..

7. Model Deployment

12. Handling machine data from a training. dataset involves several steps to ensure the data is clean, relevant & suitable. Here's a concise approach:

1. Data Cleaning
    - Removing Duplicates.
    - Handling missing values.

2. Data Transformation
   - Normalization
   - Encoding
   - Feature selection

3. Data Splitting
   - Trainning & Testing data.

4. Feature Engineering.

5. Data Augmentation & validation.

14. We don't require additional dataset for testing purpose & evaluating the accuracy.

15.

16. For migrating a large dataset to Oracle, ELT is generally more suitable due to its efficiency with large data volumes. and the flexibility of transformations within Oracle.

›ELT(Extract, Load , Transform)

- Process:

  Extract: Data is extracted from the source system.

  Load : Loaded into the target system (Oracle)

  Transform: Transform are applied directly within the target system using.

  SQL or other tools.

17. Apache Spark can be used for data integrity, data manipulation & large database management. As it is the worlds fastest & most advanced big data handling tool.

2nd Phase

1. Feature Engineering : It is the process of transforming raw data to improve the importance of machine learning.

models by creating, selecting & transforming features.

Apply Regression in Feature Modeling:

i) Interacting polynomial featu to captum non linear relation ship.

ii) Using regularisation technique like Lasso/Rodge to select important features.

iii) Normalize featu or encode categorical variables.

iv) Handling missing valu with mean, media, mode.

2 ~~Reg~~ Regularization is used in ML & statistics to prevent overfitting & improve the generalization of a model.

It add penalty to the loss function,

encouraging simpler models by reducing the magnitude of models co-efficient.

3. Yes! Linear Regression can be used for time series analysis.

4. The sum of the residuals in a linear regression should be ideally close to zero. This indicates that the model is balanced and captures the variability in the data well.

5. Multi collinearity can inflate the variance of the coefficient estimates & make them unstable & less reliable.

6. The normal form of linear regression is $y = mx + c$

7. If the beta values for a certain variable vary widely across different subsets of data, it indicates instability in the regression model.

8. The issue is likely due to perfect multicollinearity in the dataset, making it impossible for the ordinary least square (OLS) method to provide unique solution.

9. The residuals vs fitted value curve helps in diagnosing the linearity assumption in linear regression. If the points on the plot are randomly scattered around the horizontal axis without any clear pattern, it suggests that the linearity assumption holds.

10. Sermenticity refers to the situation in regression analysis where the variance of the errors is not constant across all levels of the independent variables.

To overcome it:

> We need to transform the dependent variable.

> Use weighted least squares regression.

11. Linear regression is suitable for data when there exists Linearity, Homoscedasticity, Independence, Normality among the dataset variables.

12. Hypothesis testing in linear regression evaluates the significance of relationships between variables using tests like the t-test

or f-test, determining the validity of model's predictors & the model itself.