DSBDA Viva Preparation (Simplified & Professional)

Experiment 5: Logistic Regression - Predicting Purchase Behavior

This experiment helps us predict whether someone will buy a product based on their Age and Estimated Salary.

Key Steps:

- We use 'Age' and 'Salary' as input features.
- The label is 'Purchased' (0 = No, 1 = Yes).
- The data is split into training and testing sets.
- We apply scaling to normalize values.
- A Logistic Regression model is trained.
- We check the model's performance using Accuracy, Confusion Matrix, and Classification Report.

How to Explain:

Logistic Regression is used when we need to predict a binary output (like Yes or No). It's a supervised learning technique used widely in marketing and customer behavior prediction.

Experiment 6: Naive Bayes - Predicting Flower Species

This experiment is about identifying flower species using petal and sepal measurements from the Iris dataset.

Key Steps:

- Load the Iris dataset.
- Use features like petal length/width.
- Split data into training and testing sets.
- Train a Gaussian Naive Bayes model.
- Evaluate the model using a classification report.

How to Explain:

Naive Bayes works on probability and assumes features are independent. It is simple and effective, especially for classification tasks.

DSBDA Viva Preparation (Simplified & Professional)

Experiment 7: Natural Language Processing - Text Preprocessing

In this experiment, we clean up and process a paragraph to extract meaningful words for analysis.

Key Steps:

- Break the paragraph into words (Tokenization).
- Remove common unhelpful words (Stopwords).
- Convert words to base form (Stemming & Lemmatization).
- Identify each word's role (POS Tagging).

How to Explain:

Text data is messy. We clean and structure it so machines can understand and process language, such as in chatbots or document analysis.

Experiment 8: Data Visualization - Titanic Dataset

We visualize the Titanic dataset to understand relationships like age vs fare and survival based on gender or class.

Key Steps:

- Load Titanic dataset using Seaborn.
- Use jointplot for age vs fare distribution.
- Use countplot and barplot for category comparisons.
- Use KDE and histplot for continuous feature visualization.

How to Explain:

Visualizations help us understand patterns in data, such as which groups paid more or who had higher survival chances.

Experiment 10: Data Visualization - Iris Dataset

DSBDA Viva Preparation (Simplified & Professional)

This experiment helps us understand the relationships between petal and sepal features in Iris flowers using plots.

Key Steps:

- Load the Iris dataset.
- Use pairplot to view multiple relationships.
- Use histplot and KDE to see how values are spread.
- Use boxplots to detect outliers.

How to Explain:

This helps us visually confirm which features best separate the flower types and understand the spread and range of values.

Glossary - Simple Definitions

Glossary of Simple Terms:

- Dataset: A collection of data in table format (like an Excel sheet).
- Feature: A column in the dataset used to make predictions.
- Label: The output we want to predict (also called 'target').
- Classification: Predicting categories (like Yes/No, types of flowers).
- Logistic Regression: A method used to predict Yes or No answers.
- Naive Bayes: A method that predicts using probability and assumes features are independent.
- Tokenization: Splitting text into words or sentences.
- Stopwords: Common words (like 'is', 'the', 'in') that we remove to focus on important words.
- Stemming: Cutting words to their root form (e.g., 'running' to 'run').
- Lemmatization: Like stemming, but more accurate and grammar-based.
- POS Tagging: Identifying if a word is a noun, verb, adjective, etc.
- Jointplot: A graph showing how two things relate (scatter + curve).
- KDE Plot: A smooth curve that shows how values are spread.

DSBDA Viva Preparation (Simplified & Professional)

- Boxplot: A plot that shows the range and outliers in data.
- Pairplot: Many plots showing how each feature relates to the others.
- Accuracy Score: How many predictions were correct.
- Confusion Matrix: A table that shows right and wrong model predictions.
- Precision/Recall/F1-Score: Metrics that tell how good your predictions are.