## Questions Based on Assignments :

### Assignment B-1 : Predict the price of the Uber ride

1. **What is data preprocessing?**
   Data preprocessing is the process of cleaning and transforming raw data to prepare it for analysis, ensuring it's complete and ready for machine learning.

2. **Define Outliers.**
   Outliers are data points that are significantly different from most other data, which can affect analysis accuracy.

3. **What is Linear Regression?**
   Linear Regression is a statistical method that models the relationship between two variables by fitting a straight line to the data, predicting the dependent variable based on the independent variable.

4. **What is Random Forest Algorithm?**
   Random Forest is an ensemble learning algorithm that creates multiple decision trees and combines their results to improve prediction accuracy and reduce overfitting.

5. **Explain: pandas, numpy.**

- **Pandas:** A Python library for data manipulation and analysis, providing tools to work with data in tables.
- **NumPy:** A library for numerical computing in Python, used for working with arrays and performing mathematical operations.

### Assignment B-2 : Classify the email using the binary classification method

1. **Data Preprocessing**
   Data preprocessing involves cleaning, transforming, and organizing raw data to make it suitable for analysis or machine learning models.

2. **Binary Classification**
   Binary classification is a type of classification where there are only two possible outcomes, such as "yes" or "no," "true" or "false."

**Study material provided by: Vishwajeet Londhe**

## Join Community by clicking below links

Telegram Channel

https://t.me/SPPU_TE_BE_COMP

(for all engineering Resources)

SPPU Engineering & Technical UPDATES

WhatsApp channel

WhatsApp Channel

(for all Engg & tech updates)

https://whatsapp.com/channel/0029ValjFriICVfpcV9HFc3b

@SPPU_ENGINEERING_UPDATE

Insta Page

(for all Engg & tech updates)

https://www.instagram.com/sppu_engineering_update

3. **K-Nearest Neighbours (K-NN)**
   K-NN is a simple algorithm that classifies a data point based on the majority class among its closest K neighbors in the dataset.

4. **Support Vector Machine (SVM)**
   SVM is a supervised learning algorithm that finds the best boundary (hyperplane) to separate data into different classes with maximum margin.

5. **Train, Test, and Split Procedure**
   This process divides a dataset into training and testing sets, where the training set trains the model, and the test set evaluates its performance.

**Assignment B-3 : :Given a bank customer, build a neural network-based classifier that can determine whether they will leave or not in the next 6 months**

1. **Artificial Neural Network (ANN)**
   ANN is a computational model inspired by the human brain, consisting of interconnected nodes (neurons) that learn patterns in data for tasks like classification and prediction.

2. **Keras**
   Keras is a high-level neural network library in Python, built on top of TensorFlow, that simplifies building and training deep learning models.

3. **TensorFlow**
   TensorFlow is an open-source machine learning framework that supports building, training, and deploying large-scale machine learning and deep learning models.

4. **Normalization**
   Normalization is the process of scaling data to a standard range (usually 0 to 1) to ensure each feature contributes equally to model performance.

5. **Confusion Matrix**
   A confusion matrix is a table that shows the performance of a classification model by comparing actual vs. predicted values for each class, helping to evaluate accuracy, precision, and recall.

1. **Accuracy**

   Accuracy is the percentage of correct predictions out of all predictions made. It's calculated as:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Predictions}}$$

2. **Precision**

   Precision is the percentage of true positive predictions out of all positive predictions made. It shows how accurate positive predictions are:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

3. **Recall**

   Recall is the percentage of true positive predictions out of all actual positives. It shows how well the model captures actual positive cases:

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

**Assignment B-4 : Implement K-Nearest Neighbors algorithm on diabetes.csv+**

1. **What is data preprocessing, and why is it important?**
   Data preprocessing prepares raw data by cleaning, transforming, and organizing it to improve model performance and accuracy.

2. **How do you identify outliers in a dataset?**
   Outliers can be identified using methods like the Z-score, IQR (Interquartile Range), or visualizations like box plots to detect unusual data points.

3. **What is correlation, and why is it checked?**
   Correlation measures the relationship between two variables. Checking it helps identify dependencies or multicollinearity, which can impact model effectiveness.

4. **Explain the K-Nearest Neighbors (KNN) algorithm.**
   KNN classifies data points based on the class of the K closest neighbors, making predictions based on majority voting among neighbors.

5. **What is the Random Forest algorithm?**
   Random Forest is an ensemble algorithm that creates multiple decision trees and aggregates their results to enhance prediction accuracy and reduce overfitting.

6. **What is a confusion matrix, and how is it used?**
   A confusion matrix is a table showing actual vs. predicted values in classification. It helps measure performance metrics like accuracy, precision, recall, and F1-score.

7. **Define accuracy_score.**
   Accuracy score is the percentage of correct predictions out of all predictions.

8. **What is mean_squared_error, and why is it used?**
   Mean Squared Error (MSE) measures the average squared difference between predicted and actual values, evaluating model accuracy in regression tasks.

9. **Explain r2_score and its significance.**
   R-squared (r2_score) indicates the proportion of variance explained by the model. It shows how well the model fits the data, with values closer to 1 being better.

10. **What is roc_auc_score, and why is it important?**
    The ROC AUC score measures a model's ability to distinguish between classes. It is the area under the ROC curve, where a score closer to 1 indicates better classification.

11. **Describe the ROC curve and its purpose.**
    The ROC curve is a plot of True Positive Rate vs. False Positive Rate for different thresholds, illustrating the trade-off between sensitivity and specificity in classification.