TSSM's

BHIVARABAI SAWANT COLLEGE OF ENGINEERING & RESEARCH

Narhe, Pune

Department of Computer Engineering

LABORATORY MANUAL

2023-2024

Laboratory Practice-III

BE-COMPUTER ENGINEERING

SEMESTER-I

Subject Code:**410246**

TEACHING SCHEME	***EXAMINATION SCHEME***
	Practical: 50 Marks
Practical: 4 Hrs/Week	Term Work: 50 Marks

-: **Name of Faculty** :-

S D KAMBLE

Group B Machine Learning :

Sr No	Title	Pg No
1	Predict the price of the Uber ride from a given pickup point to the agreed drop-off location. Perform following tasks: 1. Pre-process the dataset. 2. Identify outliers. 3. Check the correlation. 4. Implement linear regression and random forest regression models. 5. Evaluate the models and compare their respective scores like R2, RMSE, etc. Dataset link: https://www.kaggle.com/datasets/yasserh/uber-fares-dataset	
2	Classify the email using the binary classification method. Email Spam detection has two states: a) Normal State – Not Spam, b) Abnormal State – Spam. Use K-Nearest Neighbors and Support Vector Machine for classification. Analyze their performance. Dataset link: The emails.csv dataset on the Kaggle https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv	
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. Dataset Description: The case study is from an open-source dataset from Kaggle.The dataset contains 10,000 sample points with 14 distinct features such as CustomerId, CreditScore, Geography, Gender, Age, Tenure, Balance, etc Perform following steps:	
4	Implement Gradient Descent Algorithm to find the local minima of a function. For example, find the local minima of the function $y=(x+3)^2$ starting from the point x=2.	
5	Implement K-Nearest Neighbors algorithm on diabetes.csv dataset. Compute confusion matrix, accuracy, error rate, precision and recall on the given dataset. Dataset link : https://www.kaggle.com/datasets/abdallamahgoub/diabetes	
6	Implement K-Means clustering/ hierarchical clustering on sales_data_sample.csv dataset. Determine the number of clusters using the elbow method. Dataset link : https://www.kaggle.com/datasets/kyanyoga/sample-sales-data	
7	Mini Project	

Assignment No:1

Title:

Predict the price of the Uber ride from a given pickup point to the agreed drop-off location. Perform following tasks:

1. Pre-process the dataset.

2. Identify outliers.

3. Check the correlation.

4. Implement linear regression and random forest regression models.

5. Evaluate the models and compare their respective scores like R2, RMSE, etc.

Experiment Objective:

The objective of this lab is to predict the price of an Uber ride based on a given pickup point and agreed drop-off location. We will perform data preprocessing, outlier identification, correlation analysis, implement linear regression and random forest regression models, and evaluate their performance using metrics like R2, RMSE, etc.

Dataset Description:

The project is about on world's largest taxi company Uber inc. In this project, we're looking to predict the fare for their future transactional cases. Uber delivers service to lakhs of customers daily. Now it becomes really important to manage their data properly to come up with new business ideas to get best results. Eventually, it becomes really important to estimate the fare prices accurately.

Link for Dataset:https://www.kaggle.com/datasets/yasserh/uber-fares-datase.

Tools and Libraries:

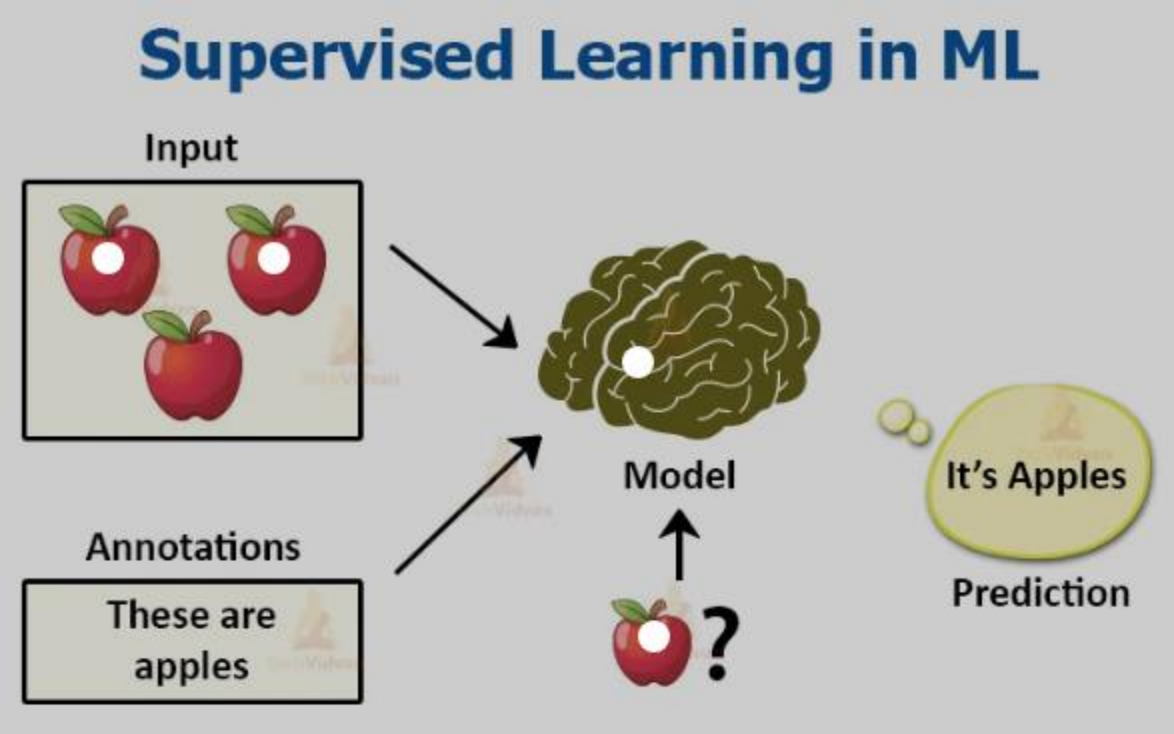
We will use Python programming language and the following libraries:

- Pandas: For data manipulation and preprocessing.
- Scikit-learn: For implementing KNN and SVM classifiers.
- Matplotlib/Seaborn: For data visualization and result analysis.

Theory:

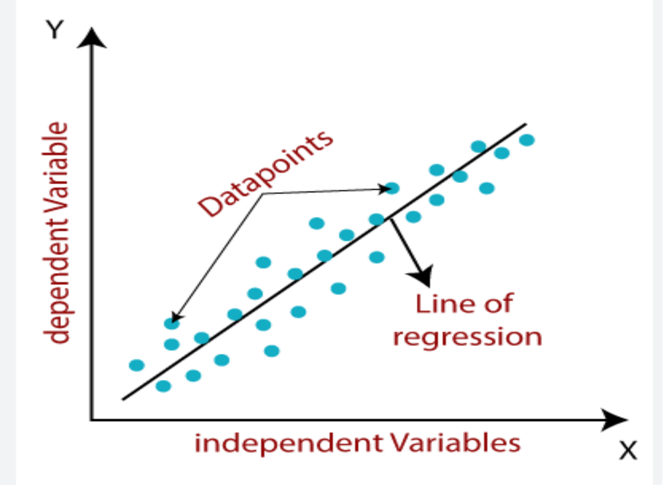
To understand and perform the practical ,you will need to have a good grasp of the following theoretical concepts in machine learning:

1. Supervised Learning: This is a type of machine learning where the model learns from labeled training data to make predictions or decisions. In this practical, we are using supervised learning to predict the price of an Uber ride based on given pickup and drop-off locations.



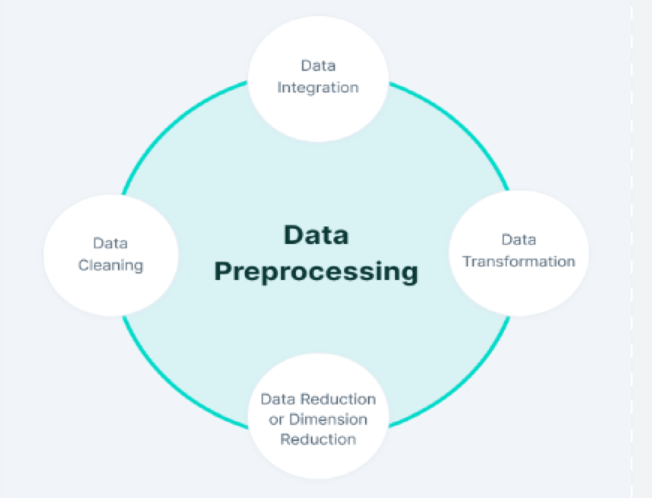
2. Regression Analysis: Regression analysis is a statistical method used to model the relationship between a dependent variable (fare_amount of the Uber ride) and one or more independent variables (pickup location, drop-off location). It helps us understand the impact of independent variables on the dependent variable and make predictions.

3. Linear Regression: Linear regression is a linear approach to modeling the relationship between the dependent variable and one or more independent variables. It assumes a linear relationship between the variables and aims to find the best-fit line that minimizes the difference between the predicted and actual values.



4. Random Forest Regression: Random forest regression is an ensemble learning method that combines multiple decision trees to make predictions. It creates a "forest" of decision trees and averages their predictions to obtain a final prediction. Random forest models are known for their ability to handle non-linear relationships and capture complex patterns in the data.

5. Data Preprocessing: Data preprocessing involves transforming and cleaning the raw data to make it suitable for analysis and modeling. This step includes handling missing values, converting categorical variables into numerical representations, and scaling or normalizing numerical features.



6. Outlier Detection: Outliers are data points that significantly deviate from the normal pattern. Outlier detection involves identifying and handling these extreme values, which can negatively impact the model's performance. Statistical methods like z-score or IQR are commonly used to detect and handle outliers.

7. Correlation Analysis: Correlation analysis helps us understand the relationship between variables. It measures the statistical association between two or more variables and indicates how they change together. Correlation analysis is useful for feature selection and identifying highly correlated features that may cause multi-colinearity issues in regression models.

8. Model Evaluation Metrics: In this practical, we use several metrics to evaluate the performance of the regression models. These metrics include R-squared (R2), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Squared Error (MSE). R2 measures the proportion of variance explained by the model, while RMSE, MAE, and MSE quantify the differences between predicted and actual values.

By understanding these theoretical concepts, you will be able to comprehend the process of building regression models, handle data preprocessing tasks, identify outliers, assess feature correlations, and evaluate model performance using appropriate metrics.

Implementation Steps:

Software required:

Anaconda with Python 3.7

Algorithm

1. Import the Required Packages
2. Read Given Dataset
3. Import the Linear Regression and Create object of it
4. Find the Accuracy of Model using Score Function
5. Predict the value using Regressor Object
6. Take input from user.
7. Calculate the value of y
8. Draw Scatter Plot

1. Pre-processing the Dataset:
   - Load the Uber ride dataset using Pandas.
   - Handle missing values using imputation or dropping rows/columns.
   - Convert categorical variables (pickup and drop-off locations) into numerical representations using one-hot encoding or label encoding.
   - Normalize the numerical features to ensure all features are on a similar scale.

2. Identifying Outliers:
    - Visualize numerical features using box plots or scatter plots to identify potential outliers.
    - Use statistical methods like z-score or IQR to detect outliers.
    - Handle outliers by either removing them or applying appropriate transformations.
3. Checking Correlation:
    - Compute the correlation matrix between numerical features.
    - Visualize the correlation matrix using a heatmap to identify highly correlated features.
    - Decide whether to drop some features to address multicollinearity.
4. Implementing Linear Regression and Random Forest Regression Models:
    - Split the dataset into training and testing sets.
    - Implement linear regression using scikit-learn.
    - Train the linear regression model on the training data.
    - Implement random forest regression using scikit-learn.
    - Train the random forest regression model on the training data.
5. Evaluating the Models and Comparing Scores:
    - Make predictions using both models on the test data.
    - Calculate R2, RMSE, MAE, and MSE for each model's predictions.
    - Compare the metrics to determine which model performs better in predicting Uber ride prices.

Conclusion:

We summarized the results of the practical, including the performance of both the linear regression and random forest regression models.