Practical-1

Machine learning basics:

In this lab, we will go through the basics of machine learning.

Topics:

What is Machine Learning?
Machine learning is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and models that enable computers to learn and make predictions or decisions without being explicitly programmed.

What are the steps involved in the collection of data?
The steps involved in the collection of data typically include identifying the purpose, planning data collection, designing data collection tools, performing data collection, validating the data, and organizing it in a structured format suitable for analysis.

What are the steps for importing data in Python using different formats such as CSV, JSON, and others?

The steps for importing data in Python using different formats are specific to each format:
- For CSV (Comma-Separated Values), you can use the csv module in Python to read the file and iterate over the rows.
- For JSON (JavaScript Object Notation), you can use the json module in Python to load the data from the JSON file using the json.load() function.
- For other data formats, such as Excel, SQL, or HDF5, specific libraries or modules exist in Python to import the data. You need to import the appropriate library/module for the specific format and use the corresponding functions or methods to import the data.

Explain the preprocessing steps involved in machine learning, including removing outliers, normalizing datasets, data encoding, and handling missing data.
Preprocessing is an essential step in machine learning.
The preprocessing steps include:
- Removing outliers: Outliers, which are extreme values, can be detected using statistical measures like Z-score or interquartile range (IQR) and then removed from the dataset.
- Normalizing datasets: Normalization involves scaling the data to a standard range to prevent certain features from dominating others. Common techniques include min-max scaling and z-score normalization.
- Data encoding: Data encoding is used to convert categorical variables into numerical representations suitable for machine learning algorithms. Techniques like one-hot encoding or label encoding can be used.
- Handling missing data: Missing data can be dealt with by either deleting the rows with missing values, imputing missing values with statistical measures like mean or median, or using advanced techniques like multiple imputation or regression imputation.

What are the different types of machine learning models, including supervised learning, unsupervised learning, and reinforcement learning?

• Supervised learning: Models learn from labelled examples and make predictions or classifications on unseen data. Examples include decision trees, random forests, support vector machines (SVM), and neural networks.
• Unsupervised learning: Models learn patterns and structures from unlabelled data. Examples include clustering algorithms (e.g., k-means, hierarchical clustering), dimensionality reduction techniques (e.g., PCA, t-SNE), and generative models (e.g., Gaussian mixture models).
• Reinforcement learning: Models learn through interactions with an environment and receive rewards or punishments based on their actions. Examples include Q-learning, deep reinforcement learning, and policy gradient methods.

b) What are some common parameters of machine learning models, such as learning rate, regularization, etc.?
Common parameters of machine learning models include learning rate, regularization (e.g., L1 or L2 regularization), number of hidden layers, activation functions, optimization algorithms, and ensemble methods. The optimal parameter values depend on the specific problem and the characteristics of the dataset.

Test-train data split: using a constant ratio, k-fold cross-validation
• When performing a test-train data split, the purpose of using a constant ratio is to divide the dataset into two portions: a training set and a test set. The constant ratio specifies the proportion of the data that will be allocated to each set. This ensures a consistent and controlled evaluation of the model's performance.
• K-fold cross-validation is a technique used to assess the performance of a model. It involves dividing the dataset into k equally-sized folds. The model is trained and evaluated k times, each time using a different fold as the test set and the remaining folds as the training set. The results are then averaged to provide a more robust evaluation of the model's performance.

Output Inference
• In the context of machine learning, output inference refers to the process of making predictions or decisions based on a trained model. Once a model has been trained on a dataset, it can be used to infer outputs or make predictions for new, unseen inputs. The model takes the input data and generates the corresponding predicted outputs based on the patterns and relationships it has learned during the training process.

Validation: different metrics - Confusion Matrix, Precision, Recall, F1-score
•**Confusion matrix** is a table that displays the true positive, true negative, false positive, and false negative predictions of a classification model. It provides a comprehensive overview of the model's performance by highlighting the accuracy and errors in predictions. It is particularly useful when dealing with imbalanced datasets.
• **Precision** is a metric that measures the accuracy of positive predictions made by a model. It is the ratio of true positives to the sum of true positives and false positives. Precision indicates how well the model correctly identifies positive samples.
• **Recall**, also known as sensitivity or true positive rate, is a metric that measures the ability of a model to correctly identify all positive samples. It is the ratio of true positives to the sum of true positives and false negatives.

• **F1-score** is a metric that combines precision and recall into a single value. It is the harmonic mean of precision and recall. F1-score provides a balanced measure of a model's performance, particularly when dealing with imbalanced datasets where precision and recall may be conflicting.