**Practical 1**

**Aim:** Implement Linear Regression (Diabetes Dataset).

**Background Information:**

**Linear Regression:**

1. Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task.
2. Regression models a target prediction value based on independent variables.
3. It is mostly used for finding out the relationship between variables and forecasting.
4. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.
5. There are many names for a regression’s dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand.
6. The independent variables can be called exogenous variables, predictor variables, or regressors.
7. Linear regression is used in many different fields, including finance, economics, and psychology, to understand and predict the behavior of a particular variable.
8. For example, in finance, linear regression might be used to understand the relationship between a company’s stock price and its earnings, or to predict the future value of a currency based on its past performance.

**Diabetes Dataset:**

1. There are several datasets available online for diabetes prediction. One such dataset is available on Kaggle. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases and contains diagnostic measurements of patients to predict whether a patient has diabetes or not.
2. Another dataset is available on CDC which provides access to the latest diabetes data and statistics through the National Diabetes Statistics Report and the Diabetes Report Card.
3. Microsoft Learn also provides a diabetes dataset which has 442 samples with 10 features, making it ideal for getting started with machine learning algorithms.

**Practical 2**

**Aim:** Implement Logistic Regression (Iris Dataset).

**Background Information:**

**Logistic Regression:**

1. Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.
2. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be of a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
3. Logistic Regression is much like Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
4. In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
5. The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

**Iris Dataset:**

1. The Iris dataset was used in R.A. Fisher's classic 1936 paper, The Use of Multiple Measurements in Taxonomic Problems, and can also be found on the UCI Machine Learning Repository.
2. It includes three iris species with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.
3. The columns in this dataset are:
   * Id
   * SepalLengthCm
   * SepalWidthCm
   * PetalLengthCm
   * PetalWidthCm
   * Species

**Practical 3**

**Aim:** Implements Multinomial Logistic Regression (Iris Dataset).

**Background Information:**

**Linear Regression:**

1. Multinomial Logistic Regression is like logistic regression but with a difference, that the target dependent variable can have more than two classes i.e., multiclass or polychotomous.
2. For example, the students can choose a major for graduation among the streams “Science”, “Arts” and “Commerce”, which is a multiclass dependent variable, and the independent variables can be marks, grade in competitive exams, Parents profile, interest etc.
3. Multinomial Logistic Regression is a classification technique that extends the logistic regression algorithm to solve multiclass possible outcome problems, given one or more independent variables.
4. This model is used to predict the probabilities of categorically dependent variable, which has two or more possible outcome classes. Whereas the logistic regression model is used when the dependent categorical variable has two outcome classes for example, students can either “Pass” or “Fail” in an exam or bank manager can either “Grant” or “Reject” the loan for a person.

**Iris Dataset:**

1. The Iris dataset was used in R.A. Fisher's classic 1936 paper, The Use of Multiple Measurements in Taxonomic Problems, and can also be found on the UCI Machine Learning Repository.
2. It includes three iris species with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.
3. The columns in this dataset are:

* Id
* SepalLengthCm
* SepalWidthCm
* PetalLengthCm
  + PetalWidthCm
  + Species

**Practical 4**

**Aim:** Implement SVM classifier (Iris Dataset).

**Background Information:**

**SVM Classifier:**

1. Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning.
2. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
3. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence algorithm is termed as Support Vector Machine.

**Iris Dataset:**

1. The Iris dataset was used in R.A. Fisher's classic 1936 paper, The Use of Multiple Measurements in Taxonomic Problems, and can also be found on the UCI Machine Learning Repository.
2. It includes three iris species with 50 samples each as well as some properties about each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.
3. The columns in this dataset are:

* Id
* SepalLengthCm
* SepalWidthCm
* PetalLengthCm
  + PetalWidthCm
  + Species

**Practical 5**

**Aim:** Train and fine-tune a Decision Tree for the Moons Dataset.

**Background Information:**

**Decision Tree:**

1. A decision tree is a decision support hierarchical model that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.
2. It is one way to display an algorithm that only contains conditional control statements.
3. Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems.
4. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules, and each leaf node represents the outcome.
5. In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
6. The decisions or the test are performed based on features of the given dataset.
7. It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.
8. It is called a decision tree because, like a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
9. To build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.
10. A decision tree simply asks a question and based on the answer (Yes/No), it further splits the tree into subtrees.
11. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal but are also a popular tool in machine learning.

**Moons Dataset:**

1. Make two interleaving half circles.
2. A simple toy dataset to visualize clustering and classification algorithms.
3. It's taken from Sklearn.

**Practical 6**

**Aim:** Train an SVM regressor on the California Housing Dataset.

**Background Information:**

**SVM Regressor:**

1. Support Vector Regression as the name suggests is a regression algorithm that supports both linear and non-linear regressions.
2. This method works on the principle of the Support Vector Machine.
3. SVR differs from SVM in the way that SVM is a classifier that is used for predicting discrete categorical labels while SVR is a regressor that is used for predicting continuous ordered variables.
4. In simple regression, the idea is to minimize the error rate while in SVR the idea is to fit the error inside a certain threshold which means, work of SVR is to approximate the best value within a given margin called ε- tube.

**California Housing Dataset:**

1. The data contains information from the 1990 California census. So, although it may not help you with predicting current housing prices like the Zillow Zestimate dataset, it does provide an accessible introductory dataset for teaching people about the basics of machine learning.
2. The data pertains to the houses found in each California district and some summary stats about them based on the 1990 census data. Be warned the data isn't cleaned so there are some preprocessing steps required!
3. The columns are as follows; their names are self-explanatory:

* longitude
* latitude
* housing\_median\_age
* total\_rooms
* total\_bedrooms
* population
* households
* median\_income
* median\_house\_value
* ocean\_proximity

**Practical 7**

**Aim:** Implement Batch Gradient Descent with early stopping for SoftMax Regression.

**Background Information:**

**Batch Gradient Descent:**

1. Batch gradient descent (BGD) is used to find the error for each point in the training set and update the model after evaluating all training examples.
2. This procedure is known as the training epoch. In simple words, it is a greedy approach where we have to sum over all examples for each update.
3. Computes gradient using the whole Training sample.
4. Slow and computationally expensive algorithm.
5. Not suggested for huge training samples.
6. Deterministic in nature.
7. Gives optimal solution given sufficient time to converge.
8. No random shuffling of points is required.
9. Can’t escape shallow local minima easily.
10. Convergence is slow.

**SoftMax Regression:**

1. SoftMax regression (or multinomial logistic regression) is a generalization of logistic regression to the case where we want to handle multiple classes in the target column.
2. In binary logistic regression, the labels were binary, that is for ith observation,
3. Softmax Regression (synonyms: Multinomial Logistic, Maximum Entropy Classifier, or just Multi-class Logistic Regression) is a generalization of logistic regression that we can use for multi-class classification (under the assumption that the classes are mutually exclusive).
4. In contrast, we use the (standard) Logistic Regression model in binary classification tasks.

**Practical 8**

**Aim:** Implement MLP for classification of handwritten digits (MNIST Dataset).

**Background Information:**

**MLPClassifier:**

1. MLPClassifier stands for Multi-layer Perceptron classifier which in the name itself connects to a Neural Network.
2. Unlike other classification algorithms such as Support Vectors or Naive Bayes Classifier, MLPClassifier relies on an underlying Neural Network to perform the task of classification.
3. One similarity though, with Scikit-Learn’s other classification algorithms is that implementing MLPClassifier takes no more effort than implementing Support Vectors or Naive Bayes or any other classifiers from Scikit-Learn.

**MNIST Dataset:**

1. The MNIST database (Modified National Institute of Standards and Technology database) is a large database of handwritten digits that is commonly used for training various image processing systems.
2. The database is also widely used for training and testing in the field of machine learning.
3. It was created by "re-mixing" the samples from NIST's original datasets.
4. The creators felt that since NIST's training dataset was taken from American Census Bureau employees, while the testing dataset was taken from American high school students, it was not well-suited for machine learning experiments.
5. Furthermore, the black and white images from NIST were normalized to fit into a 28x28 pixel bounding box and anti-aliased, which introduced grayscale levels.
6. The MNIST database contains 60,000 training images and 10,000 testing images.
7. Half of the training set and half of the test set were taken from NIST's training dataset, while the other half of the training set and the other half of the test set were taken from NIST's testing dataset.
8. The original creators of the database keep a list of some of the methods tested on it.

**Practical 9**

**Aim:** Classification of images of clothing using TensorFlow (Fashion MNIST dataset).

**Background Information:**

**Classification:**

1. The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data.
2. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups.
3. Such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. Classes can be called as targets/labels or categories.

**TensorFlow:**

1. TensorFlow is a free and open-source software library for machine learning and artificial intelligence.
2. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks.
3. TensorFlow can be used in a wide variety of programming languages, including Python, JavaScript, C++, and Java.
4. This flexibility lends itself to a range of applications in many different sectors.

**Fashion MNIST Dataset:**

1. Fashion-MNIST is a dataset of Zalando's article images—consisting of a training set of 60,000 examples and a test set of 10,000 examples.
2. Zalando intends Fashion-MNIST to serve as a direct drop-in replacement for the original MNIST dataset for benchmarking machine learning algorithms.
3. Each training and test example is assigned to one of the following labels:
   * 0 T-shirt/top
   * 1 Trouser
   * 2 Pullover
   * 3 Dress
   * 4 Coat
   * 5 Sandal
   * 6 Shirt
   * 7 Sneaker
   * 8 Bag
   * 9 Ankle boot

**Practical 10**

**Aim:** Implement Regression to predict fuel efficiency using TensorFlow (Auto MPG dataset).

**Background Information:**

**Regression:**

1. Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables.
2. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed.
3. It predicts continuous/real values such as temperature, age, salary, price, etc.

**TensorFlow:**

1. TensorFlow is a free and open-source software library for machine learning and artificial intelligence.
2. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks.
3. TensorFlow can be used in a wide variety of programming languages, including Python, JavaScript, C++, and Java.

**Auto MPG Dataset:**

1. The data is technical spec of cars. The dataset is downloaded from UCI Machine Learning Repository.
2. Number of Instances: 398
3. Number of Attributes: 9 including the class attribute
4. Attribute Information:
5. mpg: continuous
6. cylinders: multi-valued discrete
7. displacement: continuous
8. horsepower: continuous
9. weight: continuous
10. acceleration: continuous
11. model year: multi-valued discrete
12. origin: multi-valued discrete
13. car name: string (unique for each instance)
14. Missing Attribute Values: horsepower has 6 missing values