ML Task-3

Task 3.1:

Data preprocessing is very crucial before performing any operations on the actual data. It’s a technique that is used to transform raw data which has been gathered or collected from different sources. Since raw data isn’t feasible for the analysis, this process has to be carried out on every piece of significant data.

Steps:

1. The first step would be to get the actual dataset which is just data which is collected for a particular problem or event arranged and presented in a proper format. Each dataset varies on the basis of the information that it contains.
2. After the getting the actual dataset, the second step would be to import the necessary Python libraries required to clean the data. The libraries that have to be imported depend on the operations that have to be performed on the data. ‘Numpy’ is a library which is used for including any type of mathematical operation in the code. It’s significantly useful for adding large multidimensional arrays and other numerical operations as well. Another library which is extremely crucial to the actual data manipulation would be ‘Pandas’. This library is mandatory as without its functions the dataset itself cannot be imported into the program. ‘Matplotlib’ is another library which can be imported for a 2D representation of the data through different charts.
3. The next step would be to import the dataset obtained in the first step through the use of the ‘Pandas’ library. This can be done with the use of the read\_csv() function.
4. The fourth step is to handle the missing data. This is imperative because if the dataset contains any missing data, it can create a huge problem for the actual ML model that we’ll be using for handling complex patterns and making various predictions.

First of the two main ways to handle missing data is to delete a particular row. A good example of this would be using the dropna() function which removes the rows containing any null values.

Another way to do this would be through calculating the mean. This involves calculating the mean of the row or column containing the missing value and inserting it at the position of the missing data, this strategy is mostly useful for numeric data.

1. After the fourth step comes Encoding categorical data. A machine learning model completely works on math and numbers, so if there are categorical variables instead of numbers it may create trouble while building the actual model. Hence it is required that these variables get encoded into numbers.
2. The sixth step comprises of splitting the dataset into the training and the test set. Training set is a subset to train a model and the test set is a subset to test the model which was trained. Some points that have to be kept in mind while creating the test set are that the test set should be large enough to yield meaningful results and the test set which is picked shouldn’t have different characteristics as compared to the training set. This is a crucial step as by doing this, the performance of a machine learning model can be enhanced. While giving the model a training set, the output is already known, however the output is predicted in case of the test set.
3. Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. If the independent variables don’t have a similar scale, then some machine learning models wouldn’t be able to function efficiently. This method helps to scale down features in a short and similar range. Scaling also helps in improving the accuracy the model depending on the algorithm. It doesn’t have any drawbacks however its effectiveness depends on the algorithm which it is being used on.

Task 3.2

Unsupervised learning: Clustering

Unsupervised learning is a type of learning in which only the inputs are presented unlike Supervised learning where both inputs and target values are present. It depends on recognizing the regularities or predicting some information. In Clustering, there is a formation of groups or ‘clusters’ on basis of their similarities. It basically aims at forming groups of homogenous data points from a heterogenous dataset. This type of learning is highly significant as in the real world not every data we work upon has a target variable, and this particular type of data cannot be analyzed using Supervised learning algorithms. It has innumerable applications in everyday life. While identifying the similarities, the model has no prior knowledge of the dataset whatsoever.

Hard Clustering: In this type of clustering, either the data point completely belongs to a particular cluster or it does not belong to that cluster at all and belongs to a different one. Hard clustering algorithms partition the data into distinct non-overlapping clusters. There is no ambiguity in the assigning of the data point whatsoever. It’s also referred to as Crisp or Traditional clustering.

Soft Clustering: Soft clustering allows for more flexible assignments of data points. Instead of assigning a data point to a single cluster completely, it assigns a probability or degree of membership of each cluster. It allows for overlapping membership in multiple clusters. This type of clustering is also termed as Fuzzy or Probabilistic clustering.