ASSIGNMENT 3

NAME : FALAK KHAN

BATCH: Machine Learning and AI Batch A3

1. Explain the difference between data pre-processing and data wrangling:

>>>

Data pre-processing- Data preprocessing can refer to manipulation or dropping of data before it is used in order to ensure or enhance performance , and is an important step in the data mining process.

Data wrangling- also called data cleaning, data remediation refers to a variety of processes designed to transform raw data into more readily used formats.

Some examples of data wrangling include: Merging multiple data sources into a single dataset for analysis**.**

2. What is Feature Engineering and how is it different from Data Pre-Processing and Data Wrangling ?

>>> Feature engineering is the process of using domain knowledge to extract features (characteristics, properties, attributes) from raw data

Feature engineering consists of the creation of features whereas preprocessing involves cleaning the data

Whereas data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. ... This process typically includes manually converting and mapping data from one raw form into another format to allow for more convenient consumption and organization of the data.

3. What is the difference between Standardization, Normalization and Scaling ?
>>> **Standardization**: Data standardization is the process of converting data to a common format to enable users to process and analyze it.
Eg:- you can tag several tables with a particular word or field, and then use the Search field to locate all the tagged tables in your data.

Normalization: Data normalization is the organization of data to appear similar across all records and fields. It increases the cohesion of entry types leading to cleansing, lead generation, segmentation, and higher quality data.

Scaling: Scaling is a technique to standardize the independent features present in the data in a fixed range. It is essential for machine learning algorithms that calculate distances between data

4. Explain One Hot Encoding in detail.

>>> **One hot encoding-**is a process of converting categorical data variables so they can be provided to machine learning algorithms to improve predictions. One hot encoding is a crucial part of feature engineering for machine learning.

It is one method of converting data to prepare it for an algorithm and get a better prediction. With this, we convert each categorical value into a new categorical column and assign a binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector. All the values are zero, and the index is marked with a 1.

One hot encoding is useful for data that has no relationship to each other. Machine learning algorithms will read a higher number as better or more important than a lower number. In particular, one hot encoding is used for our output values, since it provides more nuanced predictions than single labels.

One hot encoding makes our training data more useful and expressive, and it can be rescaled easily .

5. How does Scaling/Normalizing data increase the performance of ML model ?

>>>The ML model can be updated to scale the target variable. Reducing the scale of the target variable will, in turn, reduce the size of the gradient used to update the weights and result in a more stable model and training process which inturn enhances the performance of the ML model.