ISE 2 (Assignment)

Arya N Kumbhar
A47 (2122000375)

GoodLuck | Page No.
Date 26 | 11 | 23.

* Data Science *

1. Explain what is mean by train/test split

- Train test split -
  These technique is used to estimate the performance of machine learning algorithms, which are used to make prediction on data not used to train the model.

- Splitting dataset is essential for an unbiased evaluation of prediction performance.
- The train & test split procedure is appropriate, when you have a very large dataset, a costly model to train, or require a good estimate model performance.
- It can be used for classification or regression problems & can be used for any supervised learning algorithm.

- Train Dataset - used to fit the machine learning model.
- Test Dataset - used to evaluate the fit machine learning model.

Process for train/test splits -

i. Taking a dataset & dividing it into two subsets.

ii. First subset is used to fit the model & reffered as training dataset.

iii. The second subset is provided as the input element to the model then predictions are made & compared to the expected values.

iv. That one is reffered as the test dataset.

GoodLuck | Page No.
Date

2. Write advantages of Numpy & matplotlib.

- Numpy -

Numerical python (Numpy) is a powerful library in python that provides support for large, multi-dimensional arrays & matrices along with mathematical functions.

- Advantages of Numpy -

i. Efficiency: Optimized for numerical computations, with C & Fortran.

ii. Array operation: Supports efficient operation on multidimensional arrays.

iii. Broadcasting:

iv. Mathematical functions: Extensive collection of math functions.

v. Provides comprehensive functions of Linear algebra operations.

vi. Random number generation.

vii. Integration.

viii. Interoperability - Easily converts to & from other python data structures.

ix. Large Community support & comprehensive documentation.

GoodLuck | Page No.
Date

- Matplotlib -
  It's a widely-used data visualization library in Python.
- Advantages of Matplotlib -
  i. Matplotlib provides a wide variety of plot types. This versatility make it suitable for various data visualization.
  ii. Publication - quality Plots.
  iii. Customization - Offers extensive customiza-tions options for plot.
  iv. Wide Apportions.
  v. It seamlessly integrates with Jupyter Notebook.
  vi. It supports 3D plots.
  vii. Matplotlib is a large ecosystem.
  viii. Active development.
  ix. Matplotlib is being open-sour-ce, It was freely available.

Date

8. Write the uses of Data Science in Health care industry.

- The efficiency of healthcare workers is increased due to the constant improvement & innovations in the industry.
- The important role of data science in health care is crucial to integration data modeling method.
- Uses of Data science in Health care -
  - i. To gather patient information dor data.
  - ii. To assess the needs or requirements of the hospital.
  - iii. To organize & classify data so that medical practitioners can use it.
  - iv. To execute data analyses using a veriety of tools.
  - v. Implement various methods to draw insights from unstructed or raw data
  - vi. Using the assistance of the development terms to create predictive models.

GoodLuck | Page No.
Date

4. Explain what is Data cleaning & munging.

- Data Cleaning :-
Data cleaning is a process of preparing & validating data before any analysis.

1. It involves fixing or removing incorrect, corrupted, incorrectly formatted, duplicate or incomplete data within dataset.

2. It plays a significant part in building a model.

3. It surely isn't fancient part of machine learning & at the same time, there aren't any hidden tricks or secrets to uncover.

4. The success or failure of project relies on proper data cleaning.

5. 'Better data beats fancier algorithms' as per the professional data scientist.

GoodLuck | Page No.
Date

1. Data Munging :-

Data Munging is the process of cleaning & transformation of data prior to use or analysis.

1. Without right tools, data munging or wrangling can be natural, time-consuming & error-prone.

2. Many organizations use tools such as Excel for data Munging.

3. But excel lacks, the sophistication & automation to make the process efficient.

4. Data munging can be a time consuming & disjointed process that stands in the way of extracting true value & potential from data.