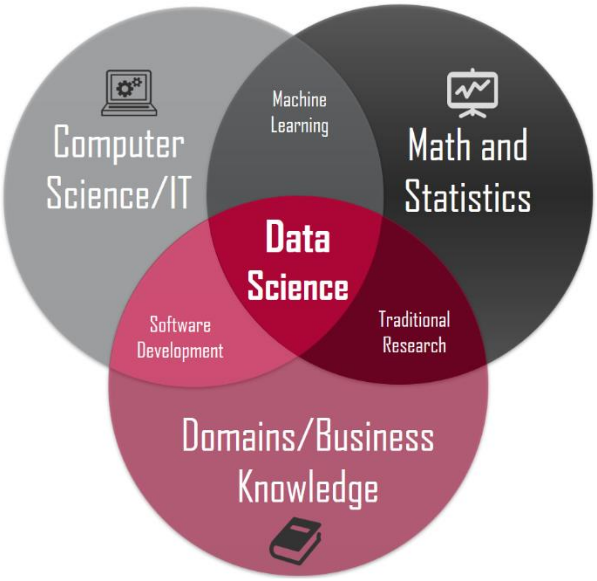
Assignment-1 Data Science and Machine Learning

Q1. What is Data Science? What are its applications?

- Data Science is the process of understanding the facts providing information from data used for decision making and business development. Technically speaking, data science combines computer analytics, business insights, and mathematical understanding.



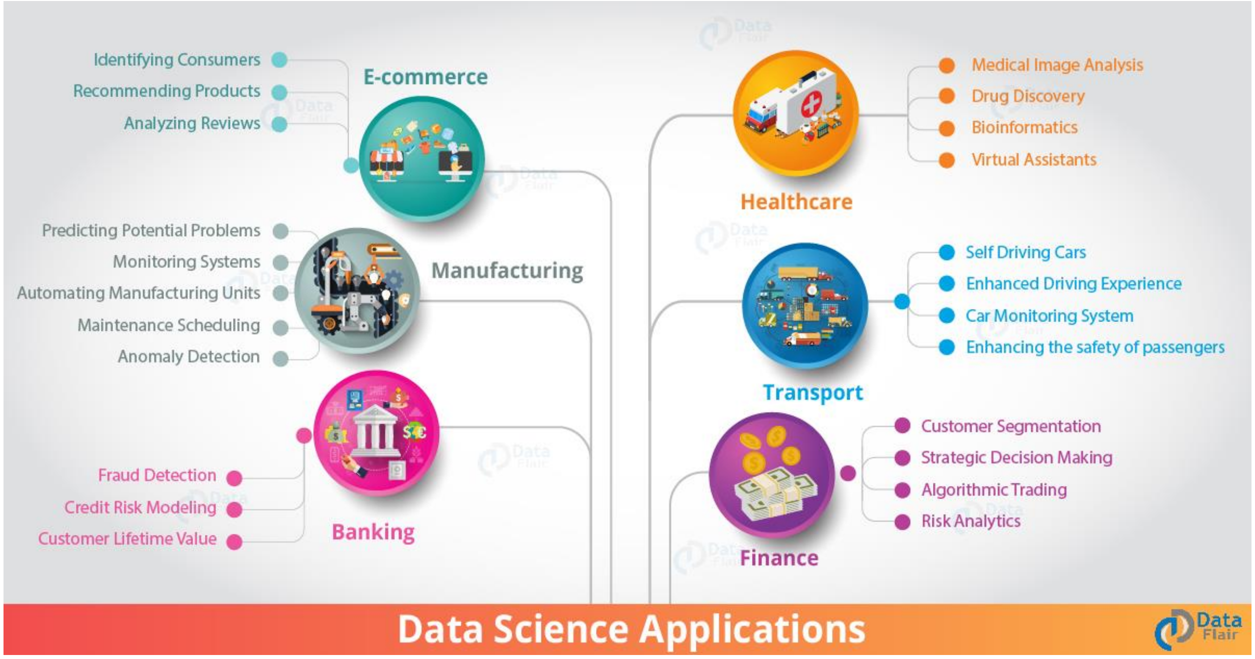
Source: Towards Data Science, Link Here

- Machine Learning, Deep Learning, Neural Networks, Artificial Intelligence, and many fields will be derived under Data Science. As the technology grows the fields under the data science also grows rapidly.
- Data Science applications are available in multiple sectors and utilize the facts from the user to provide better and user-friendly information required. Some of the areas where data science is being implemented are discussed here:
    1. Usability in social media
        - Recommendation System – by monitoring the videos user is watching, posts users are liking, products users are searching,

and tweets user is liking, the user interest-based system is created for business development.

- Spam Identification – The data science model uses the information from multiple users like which mails they are blocking and reporting, using such data the spam/junk mails will be automatically blocked for any new users.

2. Usability in healthcare
   - Hospital patients' demographics (using the patient data and appointment information)
   - Quick identification of cancer / heart-attack (using machine learning)

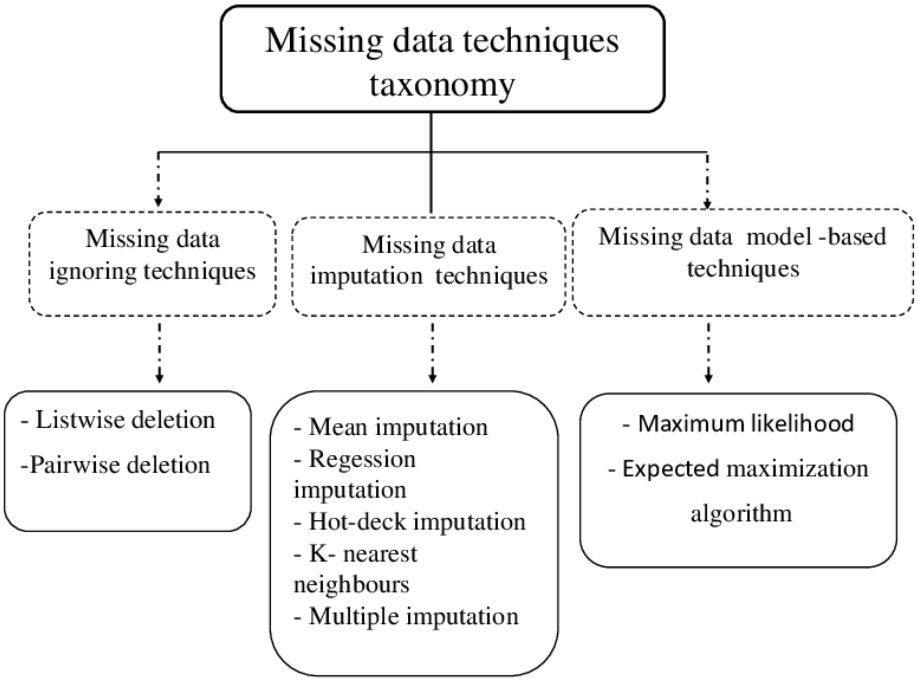


Source: Data Flair, Link Here

Q2. What are the missing values and errors in data?

- Missing Values – Data in general consists of records and features. The records count as the instances/observations (rows), whereas the features represent the data behavior (columns). The missing values are the incomplete data / missing data/wrong data/error data.
- Handling missing values - There are multiple ways to handle the missing data (Link Here). Here are some of the essential methods:
    1. **Imputation** - filling the values in the place of missing data. In this technique the empty spots in the dataset will be filled with the mean value or predicted value.
    2. **Omission** - The instances with invalid data are removed using data preprocessing methods. In this technique the invalid data like numerical in the place of categorical, will be removed using data pre-processing tools.
    3. **Analysis** - The instances with invalid data are removed using advanced algorithms/models. In this technique the invalid data will be removed and changed with algorithms.
- Missing Data types –
    1. MCAR - Missing Completely at Random - deletion is the method to handle.
    2. MAR - Missing at Random - imputation is the method to handle.
    3. MNAR - Missing Not at Random - improve/sensitize the method to handle.
- Reason for handling missing values – The missing values will increase the noise in the dataset and decreases the accuracy of the model. If the missing

values are not handled there are many problems while doing data visualizing and Hyper-parameter evaluation.



Missing data techniques taxonomy
Missing data ignoring techniques
Missing data imputation techniques
Missing data model -based techniques
- Listwise deletion
-Pairwise deletion
- Mean imputation
- Regession imputation
- Hot-deck imputation
- K- nearest neighbours
- Multiple imputation
- Maximum likelihood
- Expected maximization algorithm

Source: Research Gate, Link Here

- Experiment on missing values: The dataset used for this experiment is titanic dataset, the source is from Kaggle, and the programming language is Python. Here are the steps involved the handling the missing values from the titanic dataset:
    1. In the Python notebook, import the packages required for the data preprocessing
    2. Read the dataset from the Kaggle repository. Link for dataset here.
    3. Examine the dataset – perform some basic visualizations for understanding. Perform **describe** function to find mean, median, and other mathematical statistics.

4. Create a function to gather the missing values, calculate the count and percentage of the missing data.
5. Detecting the missing values using **missingno** package.
6. Perform the missing value handling functions – dropping columns, imputation using **sklearn** package.
7. Perform missing value handling using the KNN algorithm.

References

1. https://matteland.medium.com/what-is-data-science-c8bf9e98efd4
2. https://pub.towardsai.net/what-is-machine-learning-ml-b58162f97ec7
3. https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184
4. https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779
5. https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4
6. https://medium.com/analytics-vidhya/why-it-is-important-to-handle-missing-data-and-10-methods-to-do-it-29d32ec4e6a