Under the guidance of Internal Quality Assurance Cell (IQAC) &

In collaboration with University of Mumbai

Approved by Adhoc Board of Studies MCA, University of Mumbai

Vivekanand Education Society's Institute Of Technology

Department of MCA

Two Days Online Faculty Development Programme

On

Deep Learning Using Python

Topic : **Data Processing Techniques**

Mrs. Lifna C S (Assistant Professor @ VESIT)

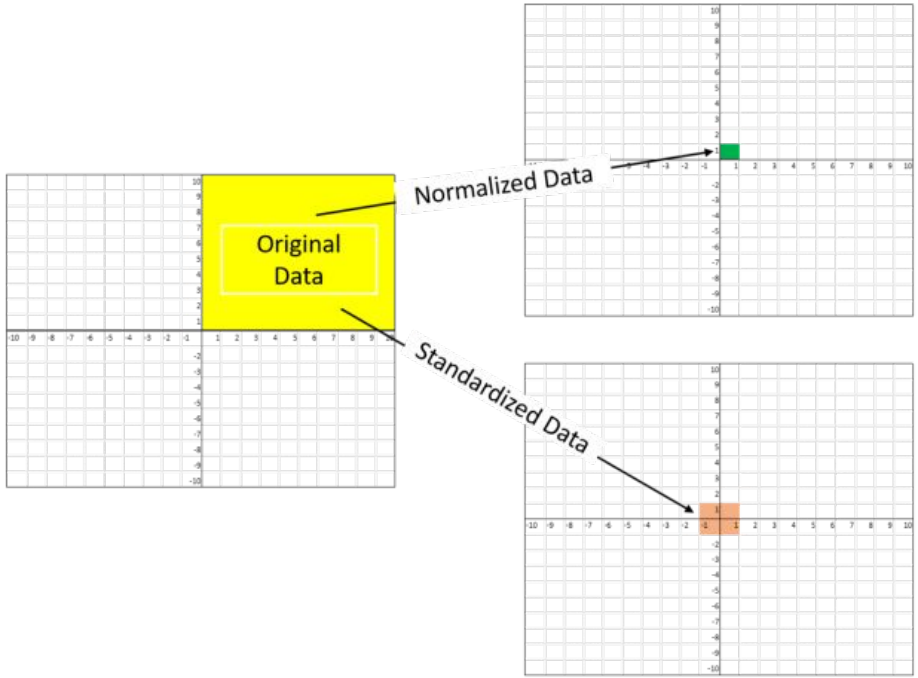
29th October 2021

Data Preprocessing Techniques

1. Data Exploration
2. Handling Numerical Data
   - Feature Scaling
   - Handling Missing Data
3. Handling Categorical Data
   - Encoding
4. Handling Textual Data
5. Data Splitting for Training & Testing

Handling Numerical Data - Feature Scaling

- **most critical step** in Data pre-processing
- most common techniques:
  - **Normalization**
  - **Standardization**



Handling Numerical Data - Missing Data

- **Three types :**
  - Missing Completely At Random (MCAR)
  - Missing At Random (MAR)
  - Missing Not At Random (MNAR)
- **Solution :**
  - Impute the data
  - Remove them (Not advisable)

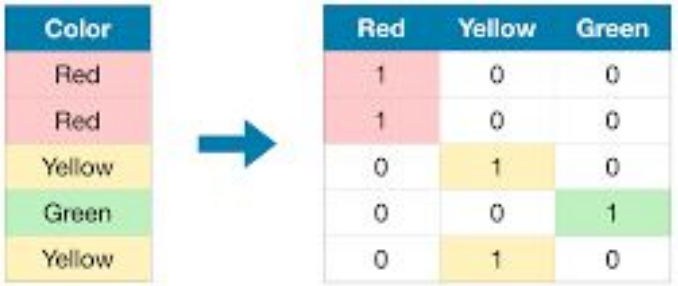
ID	Name	Age	City	Gender
1	Ram	25	Mumbai	M
2	Shyam	34	India	M
3	Shubham		Chennai	M
4	Pooja	15	Indore	F
5	Shreya		Lucknow	F
6	Yash	24	Punjab	M
7	Sakshi			
8	Abhishek			

Handling Categorical Data

Issues : Hide and mask lots of interesting information in a data set.

Methods :

- Label Encoding
- **One Hot Encoding**

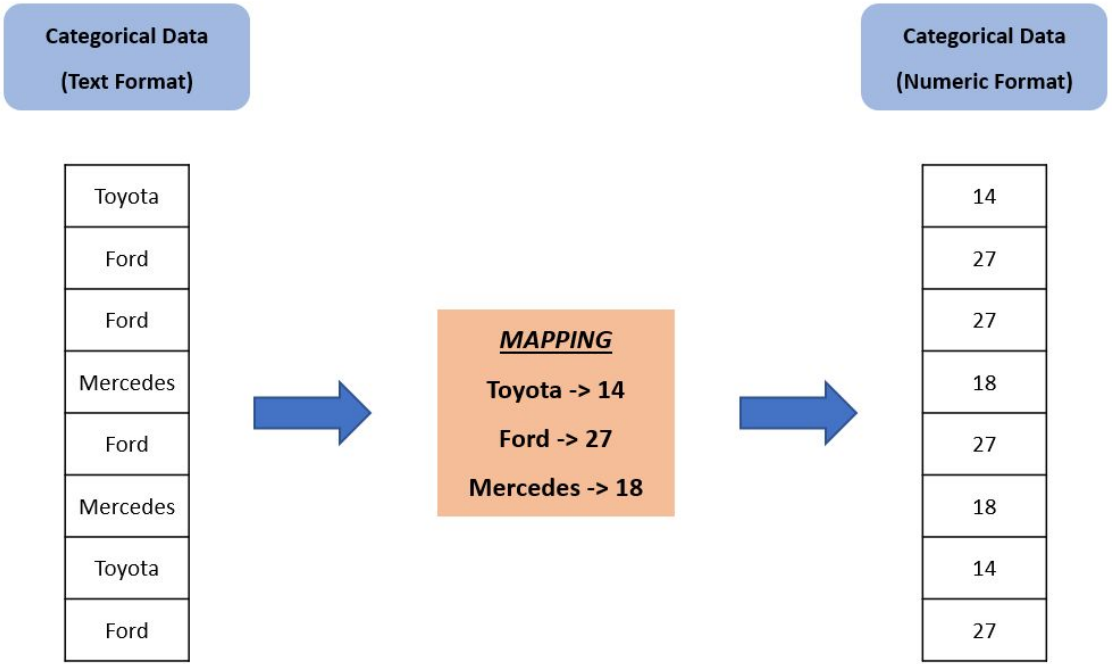


Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

Handling Categorical Data

Methods : Label Encoding



Handling Textual Data

Text normalization includes:

- converting all letters to lower or upper case
- converting numbers into words or removing numbers
- removing punctuations, accent marks and other diacritics
- removing white spaces
- expanding abbreviations
- removing stop words, sparse terms, and particular words
- text canonicalization

References

1. https://towardsdatascience.com/data-preprocessing-in-data-mining-machine-learning-79a9662e2eb
2. https://towardsdatascience.com/data-preprocessing-in-python-b52b652e37d5
3. https://towardsdatascience.com/data-pre-processing-a-step-by-step-guide-541b083912b5
4. https://www.kdnuggets.com/2020/07/easy-guide-data-preprocessing-python.html
5. https://www.section.io/engineering-education/data-preprocessing-python/
6. https://data-flair.training/blogs/python-ml-data-preprocessing/
7. https://viso.ai/deep-learning/data-preprocessing-techniques-for-machine-learning-with-python/
8. https://www.upgrad.com/blog/data-preprocessing-in-machine-learning/
9. https://www.analyticsvidhya.com/blog/2016/07/practical-guide-data-preprocessing-python-scikit-learn/
10. https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_data_preprocessing_analysis_visualization.htm
11. https://www.geeksforgeeks.org/data-preprocessing-machine-learning-python/
12. https://medium.com/@suneet.bhopal/data-preprocessing-using-python-1bfee9268fb3
13. https://www.analyticsvidhya.com/blog/2020/03/one-hot-encoding-vs-label-encoding-using-scikit-learn/
14. https://www.kaggle.com/hamelg/python-for-data-16-preparing-numeric-data
15. https://towardsdatascience.com/all-about-feature-scaling-bcc0ad75cb35
16. https://medium.com/@datamonsters/text-preprocessing-in-python-steps-tools-and-examples-bf025f872908
17. https://colab.research.google.com/github/gal-a/blog/blob/master/docs/notebooks/nlp/nltk_preprocess.ipynb#scrollTo=AdFIT_KXebxs
18. https://medium.com/machine-learning-eli5/dealing-with-categorical-data-f4c8556cbda0
19. https://www.analyticsvidhya.com/blog/2021/06/the-missing-data-understand-the-concept-behind/