

Code: BCA-3004T (For Theory) BCA-3004P (For Practical)	DSEC-I	Group-A: Elective-I Feature Engineering	1L+T:4P	3 credits (15 hours theory and 60 hours practical)
Max Marks; Theory: 100 (Int: 25; Ext: 75); Practical: 100				
<p>Course Outcomes: Upon completion of the course, students will be able to</p> <p>CO1: Understand the significance of feature engineering in the machine learning workflow and its role in enhancing model performance within the data science pipeline.</p> <p>CO2: Apply data preprocessing techniques including handling missing values, outliers, and noise, as well as scaling and normalization, to prepare raw data for modeling.</p> <p>CO3: Design and construct meaningful features through transformation techniques such as encoding, mathematical operations, and domain-specific methods; and perform feature extraction from text, image, and time-series data using advanced tools like PCA, TF-IDF, and HOG.</p> <p>CO4: Implement and evaluate feature engineering strategies using Python tools (Pandas, NumPy, Scikit-learn), apply feature selection methods (filter, wrapper, embedded), and assess their impact on model performance using metrics and cross-validation in real-world scenarios.</p> <p>Course Outcomes after Lab Programs</p> <p>CO 1. Learn Basics of Python, Pandas, NumPy, Loading and Exploring Datasets.</p> <p>CO 2. Learn data processing through data sets.</p> <p>CO 3. Learn how to Extract features from text, image, and time-series data using advanced methods like PCA, TF-IDF, and HOG</p>				
Unit	Topics			Proposed Lecture
I	<p>Introduction to Feature Engineering</p> <p>Introduction to Data and Features: Importance of Features in Machine Learning.</p> <p>Data types and features: Numerical, Categorical, Ordinal, Discrete, Continuous, Interval and Ratio.</p> <p>Basic Feature Preprocessing: Handling Missing Data, Data Cleaning, Feature Scaling, Normalization, and Transformation.</p>			2
II	<p>Feature Engineering Techniques</p> <p>Techniques for Numerical Data: Binning and Discretization, Polynomial and Interaction Features.</p> <p>Categorical Data Techniques: One Hot Encoding, Label Encoding.</p> <p>Feature extraction vs. feature selection, Steps in feature selection.</p> <p>Feature Selection Methods: Filter, Wrapper, and Hybrid.</p> <p>Feature Reduction: Introduction and application of Principal Components Analysis.</p>			2
Lab Programs	<p>The lab experiments can be implemented in Python using relevant libraries such as numpy, pandas, sklearn, nltk, matplotlib, and seaborn. Kaggle datasets, public repositories (e.g., UCI, Machine Learning etc.), or generated datasets can be used for conducting the experiments. Experiments may be conducted on numerical, image, or time-series datasets.</p> <ol style="list-style-type: none"> Handle missing values in column(s) of a dataset. For example, fill missing values with the mean/median/mode of the columns such as 'Age', 'Height', 'Weight', 'Grade' for a dataset. Clean a dataset by identifying and removing invalid data entries. For example, a dataset having columns 'Name', 'Gender' and 'Age' where 'Name' contains 'invalid data'. Scale numerical features using Min-Max normalization for a dataset with columns like 'Height', 'Weight'. Perform exploratory data analysis and visualize data distributions using histograms and box plots. 			

5. Compute and visualize the correlation matrix of a dataset with 2 or more columns.
6. Bin numerical data into discrete intervals for a dataset with a column containing numerical values.
7. Create polynomial and interaction features from numerical data in a dataset with two columns.
8. Apply logarithmic transformation to skewed numerical features in a dataset with column 'Distance'.
9. Perform one-hot encoding on categorical features in a dataset with column 'Category' containing categorical values. The distinct values in the Category feature are [Good, Better, Best] and Gender [Male, Female].
10. Preprocess text data (tokenization) for a dataset with a column 'Text'.
11. Preprocess text data (stemming) for a dataset with a column 'Text'.
12. Preprocess text data (lemmatization) for a dataset with a column 'Text'.
13. Convert text data into a Bag-of-Words representation for a dataset with a column 'Text'.
14. Apply TF-IDF transformation to text data for a column 'Text'.
15. Perform image augmentation (resizing, normalization, rotation, translation) for a set of images.
16. Perform image augmentation resizing for a set of images.
17. Perform image augmentation normalization for a set of images.
18. Perform image augmentation rotation for a set of images.
19. Perform image augmentation translation for a set of images.
20. Decompose a time series into trend, seasonal, and residual components for a dataset with a column 'TimeSeries'.
21. Perform Principal Component Analysis (PCA) on a dataset and visualize the first two principal components.

Text Books:

1. Nair, P. K. S. *Machine Learning with Python: A Practical Introduction*. Wiley India, 2020.
2. Kroese, Dirk P., et al. *Data Science and Machine Learning: Mathematical and Statistical Methods*. Pearson India, 2020.
3. Rao, R. Nageswara. *Python for Data Science*. Dreamtech Press, 2019.
4. Tripathy, B. K., and J. Anuradha. *Artificial Intelligence and Machine Learning*. Pearson India, 2020.

Reference Books:

1. Zheng, Alice, and Amanda Casari. *Feature Engineering for Machine Learning*. O'Reilly Media, 2018.
2. Gal, Soledad. *Python Feature Engineering Cookbook*. Packt Publishing, 2020.
3. Pankaj Jalote's, *Software Engineering- A Precise Approach- WILEY Precise Text Book*
4. Schach, Stephen. *Software Engineering*. 7th ed., McGraw-Hill, 2007.
5. Van Vliet, Hans. *Software Engineering: Principles and Practice*. 3rd ed., Wiley, 2008.
6. *Software Engineering (Third Edition)* -KK Agarwal, Yogesh Singh, Newage International Publishers