GUJARAT TECHNOLOGICAL UNIVERSITY

Bachelor of Engineering Syllabus

Subject Code : 3144601

Subject Name : Python for Data Science

WEF Academic Year :	2022 - 23
Semester :	4
Category of the Course :	Professional Core

Prerequisite :	Basic Knowledge of Programming and Database management systems.
Rationale :	Python is an appropriate language supporting all the features and libraries to perform Machine Learning and Data Science algorithms. This subject aims to provide the overview of the python with emphasis on various python data structures and various libraries like Pandas, NumPy, Matplotlib for performing various operations on data.

Course Scheme :

Teaching Scheme			Total Credits	Assessment Pattern and Marks				Total Marks
				Theory		Practical		
L	T	PR	C	ESE (E)	PA(M)	ESE (V)	PA (I)	
4	0	2	5	70	30	30	20	150

Course Content :

Sr. No.	Course Content	No. of Hours	% of Weightage
1	**Basics of Python** Variables and Datatypes in Python, Operators: Arithmetic, Comparison, Logical, Assignment, Bitwise, identity and membership, Strings: Slicing, String operations, List: List comprehension, List operations, Operations of Tuple, Set and Dictionary	8	15
2	**Decision Making and Looping** if statement, if...else statement, elif statement, while loop, for loop, Nested Loop, break statement, continue statement, pass statement, Use for and while loops along with useful built-in functions to iterate over and manipulate different data types.	4	10
3	**Functions and Modules** Creating User Defined Function, Parameterized function, Default argument, Keyword argument, Variable length argument, Returning value from function, global keyword, lambda function, Map, Filter and Reduce in Functional programming, Module and Package: import statement, dir() function	6	10
4	**Exception Handling and File Management** Try, except, else, finally, raise keyword, Handling multiple exceptions, Creating user defined exception, File Management: File , modes, Creating, writing and reading file, File handling functions	6	10

5	**Python Libraries for Data Science** NumPy: NumPy Array, Aggregations functions, Computation on Arrays, Comparisons, Masks, and Boolean Logic, Fancy Indexing, Sorting Arrays, NumPy's Structured Arrays Data Manipulation with Pandas: Pandas Objects, Data Indexing and Selection, Operating on Data in Pandas, Handling Missing Data, Hierarchical Indexing, Combining Datasets: Concat, Append, Merge and Join, Aggregation and Grouping Visualization with Matplotlib: 2D and 3D plots, Geographic Data with Basemap, Visualization with Seaborn	9	20
6	**Data Wrangling** Validating Data, Manipulating Categorical Variables, Dealing with Missing Data, Slicing and Dicing: Filtering and Selecting Data, Concatenating and Transforming, Aggregating Data at Any Level	6	10
7	**Exploring Data Analysis** EDA approach, Defining Descriptive Statistics for Numeric Data, Counting for Categorical Data Creating Applied Visualization for EDA, Understanding Correlation, Modifying Data Distributions **Reducing Dimensionality** : Understanding SVD, Factor Analysis and PCA, Applications of dimensionality reduction.	6	10
8	**Outliers Detection in Data** Understanding an outlier, Examining a Simple Univariate Method, Developing a Multivariate Approach Learning from Data: Linear Regression, Logistic Regression, Naïve Bayes, Learning Lazily with Nearest Neighbors, cross validation, selection and optimization	7	15
	Total	**52**	**100**

Reference Book :

1. Python for Data Science for Dummies, 2nd Edition, John Paul Mueller and Luca Massaron, Wiley Publication
2. John V Guttag. “Introduction to Computation and Programming Using Python”, Prentice Hall of India
3. Python Programming Fundamentals- A Beginner's Handbook by Nischay kumar Hegde
4. Kenneth A. Lambert, “Fundamentals of Python – First Programs”, CENGAGE Publication
5. Introduction to Python for Engineers and Scientists, By. Sandeep Nagar, Apress
6. Micro Python for the Internet of Things (A Beginner’s guide to programming with Python on microcontrollers) By. Charles Bell, Apress

Course Outcome :

After Completion of the Course, Student will able to :

No.	Course Outcomes	RBT Level*
01	Study and learn basic syntax of python programming language.	RM
02	Understand the concepts for functions, modules and exception handling.	UN
03	Explore important libraries of python related to data science.	UN
04	Apply concepts of data wrangling, EDA and Dimensionality reduction.	AP
05	Implement basic algorithms of Machine Learning related to data science.	CR

*RM: Remember, UN: Understand, AP: Apply, AN: Analyze, EL: Evaluate, CR: Create

Suggested Course Practical List :

1. Write python script for following :
   a) To understand the control structures of python.
   b) To learn different types of data structures (list, dictionary, tuples) in python.

2. Develop a program that reads a .csv dataset file using Pandas library and display the following content of the dataset.
   a) First five rows of the dataset
   b) Complete data of the dataset
   c) Summary or metadata of the dataset.

3. Develop a program to perform following operation on the dataset.
   a) To shows application of slicing and dicing over the rows and columns of the dataset.
   b) To apply split and merge operations on the dataset.

4. Develop a program that shows usage of aggregate function over the input dataset.
   a) describe b) max c) min d) mean e) median f) count g) std h) Corr

5. Develop a program that shows the various data cleaning tasks over the dataset.
   a) Identifying the null values
   b) Identifying the empty values
   c) Identifying the incorrect timestamp

6. Develop a program that shows usage of following NumPy array operations:

a) any() b) all() c) isnan() d) isinf() e) isfinite() f) isinf() g) zeros() h) isreal() i) iscomplex() j) isscalar() k) less() l) greater() m) less_equal() n) greater_equal() and vector functions: a) arrange() b) reshape() c) linspace() d) randint() e) dot()

7. Write python script using matplotlib library for following :

a) To display line plot for given values of X: [1,2,3,...,49] and values of Y (thrice of X): [3,6,9,12,...,144,147]

b) To display bar plot for given value Languages = ['Java', 'Python', 'PHP', 'JavaScript', 'C#', 'C++'] and popularity = [22.2, 17.6, 8.8, 8, 7.7, 6.7]

c) To display pie plot for given value Languages = ['Java', 'Python', 'PHP', 'JavaScript', 'C#', 'C++'] and popularity = [22.2, 17.6, 8.8, 8, 7.7, 6.7] using colors = ["#1f77b4", "#ff7f0e", "#2ca02c", "#d62728", "#9467bd", "#8c564b"]

d) To display scatter plot for 200 random points for both X and Y axis.

8. Develop a program that reads .csv file from the url: (https://github.com/chris1610/pbpython/blob/master/data/sample salesv3.xlsx?raw=true) and plot the data of the dataset stored in the .csv file.

9. Write a text classification pipeline using a custom preprocessor and CharNGramAnalyzer using data from Wikipedia articles as a training set.

a) Evaluate the performance on some held out test sets

10. Write a text classification pipeline to classify movie reviews as either positive or negative.

a) Find a good set of parameters using grid search.

b) Evaluate the performance on a held out test set.

List of Laboratory/Learning Resources Required :

- Python
- Google COLAB
- ANACONDA and Jupiter notebook

* * * * * * *