Course: Programming for Data Science
Course Code: DS222 Credit:4

Course Description:

The course DS222 —*PDS* deals with the basics of the programming languages essential to work in the field of data science and machine learning. It covers the programming constructs, data structures and useful data handling techniques in the following programming languages - Python and ANSI SQL. The enabling environment for these languages are VSCode / Python Notebooks / Jupiter Notebooks and MySQL. The process of developing an algorithm or pseudocode from a given problem statement and its translation into a working code exploiting various constructs and features of the programming language is also undertaken. Basic concepts of exception handling, function design, and object orientation are also emphasized. The use of data science specific python modules and packages such as pandas, numpy and matplotlib for exploratory data analysis (EDA) and visualisation are also covered. Basic concepts of an RDBMS such as normalization, joins and sub-queries are covered while learning the SQL syntax.

Prerequisite(s): NA

Course Objectives:

Knowledge acquired :

- syntax of the python programming language
- modularization of code and design of functions, modules and packages
- process of translating a given problem into working code
- object oriented programming and its application
- concept of exception handling and file handling
- use of pandas, numpy and matplotlib python modules and packages for data science applications
- RDBMS concepts and their application using ANSI SQL

Skills gained :

- ability to convert a given problem into python working code
- ability to employ the correct python control structure
- ability to design functions, classes and exception handling in python
- ability to modularize python code into packages and modules
- ability to apply the steps of exploratory data analysis (EDA)
- ability to use datascience specific python packages such as pandas and numpy for data handling and data manipulation
- ability to use matplotlib and seaborn python packages for data visualization
- ability to design RDBMS tables and constraints
- ability to query and update RDBMS tables using ANSI SQL.

General Competence :

- appreciate the usage and application of a particular programming language syntax
- ability to manipulate, analyse and visualize any given dataset
- ability to design and implement a basic RDBMS database model
- write bug-free code, modularised and robust code.

Approximate weightage of different components in evaluation:

Assignments/Tests	20%
Midterm Exam	40%
Final Exam	40%

Course Outline (tentative):
Getting started with Python: Setting up the python shell environment, using VSCode, Anaconda, IPython and Jupyter Notebook.

Python Data structures: python datatypes, tuple, list, set, dictionary; comprehension with list, set, dictionary; indexing, slicing and transforming data structures.

Python basics: Object model and references, Expressions, Data types, String functions, Functions and lambda functions, IPython magics, Importing and using modules, String formating, Type casting, Control structures, Exception handling, Operator precedence, Timing code blocks , Reading and writing to files.

Object Oriented Programming with Python: Concept of classes and objects, Inheritance, instance and class variables, class methods, overriding methods, special dunder methods.

Generators and Decorators of Python: Concept and use of generators and decorators with python.

Numpy: Arrays, Matrices, Arithmetic with arrays, Index- ing arrays, Linear algebra with arrays, Copying, Generating discrete and continuous distributions.

Pandas basics: Series, DataFrames, Indexing and slicing, Cre- ating, modifications to Series and Dataframes.

Data preprocessing with pandas: Missing value analysis, Duplicate data han- dling, String and date manipulation, Variable transformation - discretisation, binning, recoding, filtering, dummy variable creation, Sort, order, map, filter functions, Merging, subset- ting, sampling, reordering, reshaping datasets, Grouping and aggregate operation, Cross tabulation.

Data analysis with pandas: Use of groupby and apply functions - split-apply-combine principle, Comparison of apply, aggregate, transform, Reindexing dataframes, renaming columns, Changing datatype - categorical data, Pivot tables and cross tabulation.

Visualization with python: Matplotlib - working with the OO model (Figure, Axes, Artists, etc.), Matplotlib plotting styles - using OO model, using pyplot module, using dataframe plot function, Seaborn - faceting and advanced plots.

Regular expression with python: Use of re module.

SQL Basics: Data definition language (DDL) - create table, database, constraints, Data manipulation language (DML) - select, insert, alter and update commands, Joins - inner, outer, cartesian product, Subquery and correlated subquery.

Textbooks(s):

1. *Python for Data Analysis* by Wes McKinney, 2nd edition.
2. *Python Data Science Handbook* by Jake VanderPlas, 1st edition.

3. *SQL: The Complete Reference* by James R. Groff, Paul N. Weinberg & Andrew J. Oppel, 3rd edition.

4. *MySQL Documentation* URL: https://dev.mysql.com/doc/mysql-getting-started/en/.