| ECT332 | DATA ANALYSIS | CATEGORY | L | T | P | CREDIT |
|---|---|---|---|---|---|---|
| | | PEC | 2 | 1 | 0 | 3 |

**Preamble:** This course aims to set the foundation for students to develop new-age skills pertaining to analysis of large-scale data using modern tools.

**Prerequisite**: None

**Course Outcomes:** After the completion of the course the student will be able to

| | |
|---|---|
| **CO 1** | Read and write data to and fro spreadsheets and databases |
| **CO 2** | Work with large data as pandas data frames |
| **CO 3** | Perform PCA and cluster analysis on data frames |
| **CO 4** | Perform Bayesian analysis on data frames. |
| **CO 5** | Apply machine learning in data analysis problems |
| **CO 6** | Apply methods in high performance computing for data analysis |

**Mapping of course outcomes with program outcomes**

| | PO 1 | PO 2 | PO 3 | PO 4 | PO 5 | PO 6 | PO 7 | PO 8 | PO 9 | PO 10 | PO 11 | PO 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CO 1** | 3 | 3 | | | 3 | | | | | | | 2 |
| **CO 2** | 3 | 3 | 2 | 3 | 3 | | | | | | | |
| **CO 3** | 3 | 3 | 2 | 3 | 3 | 2 | | | | | | 2 |
| **CO 4** | 3 | 3 | 2 | 3 | 3 | 2 | | | | | | 2 |
| **CO 5** | 3 | 3 | 2 | 3 | 3 | 2 | | | | | | 2 |
| **CO 6** | 3 | 3 | 2 | 3 | 3 | 2 | | | | | | 2 |

**Assessment Pattern**

| Bloom's Category | Continuous Assessment Tests | | End Semester Examination |
|---|---|---|---|
| | 1 | 2 | |
| Remember | 10 | 10 | 20 |
| Understand | 30 | 30 | 60 |
| Apply | 10 | 10 | 20 |
| Analyse | | | |
| Evaluate | | | |
| Create | | | |

**Mark distribution**

| Total Marks | CIE | ESE | ESE Duration |
|---|---|---|---|
| 150 | 50 | 100 | 3 hours |

**Continuous Internal Evaluation Pattern:**

| | |
|---|---|
| Attendance | : 10 marks |
| Continuous Assessment Test (2 numbers) | : 25 marks |
| Assignment/Quiz/Course project | : 15 marks |

**End Semester Examination Pattern:** There will be two parts; Part A and Part B. Part A contain 10 questions with 2 questions from each module, having 3 marks for each question. Students should answer all questions. Part B contains 2 questions from each module of which student should answer any one. Each question can have maximum 2 sub-divisions and carry 14 marks.

**Course Level Assessment Questions**

**Course Outcome 1 (CO1): Read and write data to and fro spreadsheets and databases**

1. Write Python code to read an .xls file using xlrd module. Sve it as a different .xlsx file using openpyxl.
2. Write Python code to read mongodb data base.

**Course Outcome 2 (CO2): Work with pandas dataframes**

1. Write Python code read a table in a pdf file as a pandas dataframe.
2. Write Python code to create a pandas dataframe. Pickle this data and store it. Write another Python code to retrive the data from the pickle.

**Course Outcome 3 (CO3): PCA and Cluster Analysis**

1. Write Python code to perform PCA on a pandas dataframe. Write code to create a scree plot.
2. Write Python code to do K-means clustering.

**Course Outcome 4 (CO4): Bayesian Analysis on Dataframes**

1. Write Python code to compute the posterior probability of a data set with Pymc3
2. Write a python code to evaluate the statistical correlation between variables in 5X5 random data set.

**Course Outcome 5 (CO5): Machine learning in Data Analysis**

1. Write python code to use Keras for training a CNN
2. Write Python code to read an RGB image and convert to gray scale and write the grayscale image in .jpg format.

**Course Outcome 5 (CO6): High Performance Computing Methods in Data analysis**

1. Write Python code to use numexpr for faster parallel computation
2. Write Python code with Ipython-parallel to perform parallel computing with 4 cores.

## SYLLABUS

**Module 1: Overview of Data Analysis and Python**

Numpy and Scipy Python modules for data analysis. Reading and processing spreadsheets and csv files with Python using xlrd, xlwt and openpyxl. Data visualization with Matplotlib. Two dimensional charts and plots. Scatter plots with matplotlib. Three dimensional visualization using Mayavi module. Reading data from sql and mongodb databases with Python.

**Module 2: Big Data Arrays with Pandas**

Familiarization of the python pandas. Reading and writing pandas dataframes. Reading rows and columns from pandas dataframe. Handling NaN values. Reading and writing .txt, .csv, .pdf, .html and json files with pandas. Merging, concatenating and grouping of data frames. Use of pivot tables. Pickling of data frames in Python.

**Module 3: PCA and Cluster Analysis**

Singular value decomposition of a matrix/array. Eigen values and eigen vectors. Principal component analysis of a data frame**.** Scree plot. Dimensioanality reduction with PCA. Loadings for principal components. Case study with Python. Cluster analysis. Hierarchial and K-means clustering. Interpretation of dendrograms.

**Module 4: Statistical Data Analysis**

Hypothesis testing. Bayesian analysis. Meaning of prior, posterior and likelyhood functions. Use of pymc3 module to compute the posterior probability. MAP Estimation. Credible interval, conjugate distributions. Contingency table and chi square test. Kernel density estimation.

**Module 5: Machine Learning**

Supervised and unsupervised learning. Use of scikit-learn. Regression using scikit-learn. Deep learning with convolutional newural networks. Structure of CNN. Use of Keras and Tensorflow. Machine learning with pytorch. Reading and writing images with openCV. Case study of character recognition with MNIST dataset. High performance computing for machine learning. Use of numba, jit and numexpr for faster Python code. Use of Ipython-parallel.

**Text Books and References**


1. “Python Data Analytics”, Fabio Nelli, Apress.
2. “Data Analysis from Scratch with Python”, Peters Morgan, AI Sciences.
3. “Python for Data Analysis”, Wes McKinny, O’Reilly.
4. “Ipython Interactive Computing and Visualization Cookbook”, Cyrille Rossant , PACKT Open Source Publishing
5. “Deep Learning with Python”, Francois Chollet, Manning

**Course Contents and Lecture Schedule**

| No | Topic | No. of Lectures |
|---|---|---|
| **1** | **Overview of Data Analysis and Python** | |
| 1.1 | Numpy and Scipy Python modules for data analysis. | 2 |
| 1.2 | Reading and processing spreadsheets and csv files with Python using xlrd, xlwt and openpyxl. | 2 |
| 1.3 | Data visualization with Matplotlib. Two dimensional charts and plots. Scatter plots with matplotlib. Three dimensional visualization using Mayavi module. | 2 |
| 1.4 | Reading data from sql and mongodb databases with Python | 2 |
| **2** | **Big Data Arrays with Pandas** | |
| 2.1 | Intro. To Python pandas | 1 |
| 2.2 | Reading and writing of data as pandas dataframes. Separating header, columns row etc and other manipulations | 3 |
| 2.3 | Reading data from different kind of files, Merging, concatenating and grouping of data frames. Use of pivot tables. Pickilng | 3 |
| **3** | **PCA and Cluster Analysis** | |
| 3.1 | Singular value decomposition of a matrix/array. Eigen values and eigen vectors. | 1 |
| 3.2 | PCA,Scree plot. Dimensioanality reduction with PCA. Loadings for principal components. Case study with Python. Cluster analysis. | 3 |
| 3.3 | Cluster analysis, dendrograms | 2 |
| **4** | **Statistical Data Analysis** | |
| 4.1 | Hypothesis testing. Bayesian analysis. Meaning of prior, posterior and likelyhood functions. Use of pymc3 module to compute the posterior probability. | 3 |
| 4.2 | MAP Estimation. Credible interval, conjugate distributions .Contingency table and chi square test. Kernel density estimation. | 3 |
| 4.3 | Contingency table and chi square test. Kernel density estimation. | 3 |
| **5** | **Machine Learning** | |
| 5.1 | Supervised and unsupervised learning. Use of scikit-learn. Regression | 2 |
| | using scikit-learn. | |
| 5.2 | Deep learning with convolutional newural networks. Structure of CNN. | 2 |
| 5.3 | Use of Keras and Tensorflow. Machine learning with pytorch. Case study of character recognition with MNIST dataset. | 3 |
| 5.4 | High performance computing for machine learning. Use of numba, jit and numexpr for faster Python code. Use of Ipython-parallel. | 2 |

**Simulation Assignments**

1. Download the iris data set and read into a pandas data frame. Extract the heaser and replace with a new header. Extract colums and rows. Extract pivot tables. Filter the data based on the labels. Store a pivot table as a pickle and retrieve it.
2. For the same data sert, perform principal component analysis. Observe the scree plot. Identify the principal components. Obtain a low dimensional data, with only the principal components and compute the mean square error between the original data and the approximated one. Compute the loadings for the principal components.
3. For the same data, perform hierarchical and K-means clustering with Python codes. Obtain dendrograms in each case and appreciate the clusters.
4. Download the MNIST letter data set. Construct a CNN network with appropriate layers using Keras and Tensorflow. Train the CNN with the MNIST data set. Appreciate the selection and use of training, test and cross-validation data sets. Save the model and weights and use the model to identify letter images. You may use openCV for reading images.
5. Write a Python script to generate alphanumeric images (26 upper case, 26 lowercase and 10 numbers each 12 point in size) of say 16X16 dimension out of windows .ttf files. Create 62 folders each containg a data set of every alphanumeric character. Create a new CNN with Keras and Tensorflow. Create a cross validation data set by taking 10 images out of every 62 folder. Use 80% of the total data for training and 20% for testing the CNN. Use an HPCC like system to train the model and save the model and weight. Test this model to recognize letter images. You may use openCV for reading images.
6. Repeat assignment 4 using pytorch instead of Keras
7. Repeat assignment 5 using pytorch instead of Keras

**Model Question Paper**

A P J Abdul Kalam Technological University

Sixth Semester B Tech Degree Examination

Course: ECT 332 Data Analysis

Time: 3 Hrs Max. Marks: 100

PART A

*Answer All Questions*

| | | | |
|---|---|---|---|
| 1 | Create a two dimensional array of real numbers using numpy. Write Python code to pickle this data. | (3) | $K_3$ |
| 2 | Write Python code to import mayavi module and perform 3-D visualization of $x^2 + y^2 + z^2 = 1$ | (3) | $K_3$ |
| 3 | Write Python code to generate a 5 × 5 pandas data frame of random numbers. Add a header to this dataframe. | (3) | $K_3$ |
| 4 | Write Python code to concatenate two dataframes of same num- ber of columns. | (3) | $K_3$ |
| 5 | Write the expression for the singular value decomposition of a matrix $A$ | (3) | $K_3$ |
| 6 | Explain how principal components are isolated using scree plot. | (3) | $K_1$ |
| 7 | State Bayes theorem and explain the significance of the terms prior, likelyhood and posterior. | (3) | $K_1$ |
| 8 | Write Python code with pymc3 to realize a Bernoulli trial with $p(head) = 0.2$ | (3) | $K_3$ |
| 9 | Give the structure a convolutional neural network | (3) | $K_1$ |
| 10 | Compare supervised and unsupervised learning | (3) | $K_1$ |

PART B

*Answer one question from each module. Each question carries 14 mark.*

Module I

| | | | |
|---|---|---|---|
| 11(A) | Write Python code to read a spreadsheet in .xls format a text file in .csv format and put these data into numpy arrays. in both cases, plot the second column against the first column using matplotlib | (8) | $K_3$ |
| 11(B) | Write Python code to read tables from sql and mongodb databases. | (6) | $K_3$ |

OR

| | | | |
|---|---|---|---|
| 12(A) | Write Python code to create a normally distributed 5 × 5 random array and convert it into a matrix. Write code to | (8) | $K_3$ |

compute its inverse and transpose.

| | | | |
|---|---|---|---|
| 12(B) | Write code to read files in .xlsx format using openpyxl | (6) | $K_3$ |

Module II

| | | | |
|---|---|---|---|
| 13(A) | Write Python code to import a table in .xls format into a data frame. Remove all NaN values. | (6) | $K_3$ |
| 13(B) | Write Python code to generate 10 data frames of size 5 × 5 of random numbers and use a *for loop* to concatenate them. Pickle the concatenated dataframe and store it. Write another code to retrieve the dataframe from the pickle. | (8) | $K_3$ |

OR

| | | | |
|---|---|---|---|
| 14(A) | Write Python code to read in a table from a pdf file into a pandas dataframe. Write code to remove the first two columns and write the rest of the dataframe as a json file. | (8) | $K_3$ |
| 14(B) | Explain the term pivot table. Create a pivot table from the above dataframe | (6) | $K_3$ |

Module III

| | | | |
|---|---|---|---|
| 15 | Write Python code to read in table in .xls format, perform PCA analysis on it and produce the scree plot and loadings for the principal components. | (14) | $K_3$ |

OR

| | | | |
|---|---|---|---|
| 16. | Write Python code to perform hierarchial cluster analysis on a pandas dataframe. Explain how dendrograms can be used to classify data. | (14) | $K_3$ |

Module IV

| | | | |
|---|---|---|---|
| 17(A) | Assume that you have a dataset with 57 data points of Gaussian distribution with a mean of 4 and standard deviation of 0.5. Using PyMC3, write Python code to compute:<br>• The posterior distribution<br>• The prior distribution<br>• The posterior predictive distribution | (8) | $K_3$ |

.

| | | | |
|---|---|---|---|
| 17(B) | Write a python code to find the Bayesian credible interval in the above question. How is it different from confidence interval. | (6) | $K_3$ |

OR

| | | | |
|---|---|---|---|
| 18(A) | Write a python code to evaluate the statistical correlation between variables in 10 × 10 random data set. | (8) | $K_3$ |
| 18(B) | Compute the conjugate of the logarithmic function $f(x) = lnx,\ x > 0.$ | (6) | $K_3$ |

Module V

| | | | |
|---|---|---|---|
| 19(A) | Explain the use of numba and numexpr in fatser Python execution with examples | (8) | $K_3$ |
| 19(B) | Explain the use of Keras as a frontend for Tensorflow with Python codes | (6) | $K_3$ |

OR

| | | | |
|---|---|---|---|
| 20(A) | Explain the use of Ipython-parallel in parallel execution of Python code with examples. | (8) | $K_3$ |
| 20(B) | Explain with Python codes how openCV is used to read and writte images. | (8) | $K_3$ |