ADL331	AI & DATA SCIENCE LAB	CATEGORY	L	T	P	Credit	Year of Introduction
		PCC	0	0	3	2	2020

Preamble: The course enables the learners to get hands-on experience in AI and data science using Python programming. It covers implementation of various predictive and descriptive analysis measures, supervised learning algorithms (such as linear regression, logistic regression, decision trees, Bayesian learning and Naive Bayes algorithm) and unsupervised learning algorithms (such as basic clustering algorithms). This helps the learners to develop, implement algorithms and evaluate its performance for real world data.

Prerequisite: Fundamentals of programming, python programming fundamentals, Machine learning.

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

CO#	Course Outcomes
CO1	Implement various predictive and descriptive analysis measures using Python. Use various packages and libraries in Python for data handling. **(Cognitive Knowledge Level: Apply)**
CO2	Implement different Regression methods such as Linear and Logistic regression to interpret the given dataset. **(Cognitive Knowledge Level: Apply)**
CO3	Implement various supervised learning models like k-Nearest Neighbour, Support Vector Machine, Naïve Bayesian Classifier and Decision Tree algorithms. **(Cognitive Knowledge Level: Apply)**
CO4	Implement mathematical optimization method like the Hill Climbing algorithm and Deep Learning method like Convolutional Neural Network algorithm. **(Cognitive Knowledge Level: Apply)**
CO5	Implement different methods (like Correlation and Covariance) to determine the dependence between features in the dataset and apply dimensionality reduction techniques. **(Cognitive Knowledge Level: Apply)**

Mapping of course outcomes with program outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12
CO1	✓	✓	✓	✓				✓				✓
CO2	✓	✓	✓	✓				✓				✓
CO3	✓	✓	✓	✓		✓		✓				✓
CO4	✓	✓	✓	✓		✓		✓				✓
CO5	✓	✓	✓	✓		✓		✓				✓

Abstract POs defined by National Board of Accreditation			
PO#	**Broad PO**	**PO#**	**Broad PO**
PO1	Engineering Knowledge	**PO7**	Environment and Sustainability
PO2	Problem Analysis	**PO8**	Ethics
PO3	Design/Development of solutions	**PO9**	Individual and teamwork
PO4	Conduct investigations of complex problems	**PO10**	Communication
PO5	Modern tool usage	**PO11**	Project Management and Finance
PO6	The Engineer and Society	**PO12**	Lifelong learning

Assessment Pattern

Bloom's Category	Continuous Assessment Test (Internal Exam) Marks in percentage	End Semester Examination Marks in percentage
Remember	**20**	**20**
Understand	**20**	**20**
Apply	**60**	**60**
Analyze		
Evaluate		
Create		

Mark Distribution

Total Marks	CIE Marks	ESE Marks	ESE Duration
150	75	75	3 hours

Continuous Internal Evaluation Pattern:

Attendance : **15 marks**
Continuous Evaluation in Lab : **30 marks**
Continuous Assessment Test : **15 marks**
Viva voce : **15 marks**

Internal Examination Pattern:

The internal examination shall be conducted for 100 marks, which will be converted to out of 15, while calculating internal evaluation marks. The marks will be distributed as, Algorithm - 30 marks, Program - 20 marks, Output - 20 marks and Viva - 30 marks.

End Semester Examination Pattern:

The end semester examination will be conducted for a total of 75 marks and shall be distributed as, Algorithm - 30 marks, Program - 20 marks, Output - 20 marks and Viva- 30 marks.

Operating System to Use in Lab : Linux / Windows

Programming Language to Use in Lab : Python

Fair Lab Record:

All the students attending the AI & Data Science Lab should have a fair record. Every experiment conducted in the lab should be noted in the fair record. For every experiment, in the fair record, the right hand page should contain experiment heading, experiment number, date of experiment, aim of the experiment, procedure/algorithm followed, other such details of the experiment and final result. The left hand page should contain a print out of the respective code with sample input and corresponding output obtained. All the experiments noted in the fair record should be verified by the faculty regularly. The fair record, properly certified by the faculty, should be produced during the time of end semester examination for the verification by the examiners.

Syllabus

***Mandatory**

1. Implement a program to perform operations like mean, median, mode, standard deviation, percentile and various data distributions.
2. Review of python programming, Matrix operations, Programs using matplotlib / plotly / bokeh / seaborn for data visualisation and programs to handle data using pandas*
3. Try to open a csv file and sort the content with respect to one column using python.
4. Implement a program to perform linear regression for a dataset that prevails in csv format*
5. Implement a program to perform logistic regression to classify a dataset. Print feature importance after building model*
6. Implement k-Nearest Neighbour algorithm to classify any dataset. Print both correct and wrong predictions. ML library classes can be used for this problem. Assume K=3.*
7. Write a program to construct a Support Vector Machine considering medical data. Use this model to demonstrate the diagnosis of heart patients using the standard Heart Disease Data Set*
8. Assuming a set of documents that need to be classified, use the naïve Bayesian Classifier model to perform this task. Calculate the accuracy, precision, and recall for your data set*
9. Assuming a set of data that need to be classified, use a decision tree model to perform this task. Preferably use any dataset like medical or others to evaluate the accuracy.*
10. Implement a program to perform Hill climbing algorithm.*
11. Implement convolutional neural network to classify images from any standard dataset in the public domain using Keras framework. Reading and writing different types of dataset.
12. Write a program to find Correlation and Covariance between different features of a dataset in csv format.*
13. Write a program to implement feature reduction using PCA. Calculate the covariance between features to find the optimal number of PCA components.*

Practice Questions

1. Write a Python script to generate a list of random numbers and find their mean and standard deviation.
2. Consider the river temperature data available at https://catalogue.ceh.ac.uk/documents/b8a985f5-30b5-4234-9a62-03de60bf31f7. Create a Python script to select only the data from "Swale at Catterick Bridge" location, and find the mean temperature and median dissolved oxygen. Also plot a histogram showing the distribution of temperature over the time period of study.

3. Consider the river temperature data available at https://catalogue.ceh.ac.uk/documents/b8a985f5-30b5-4234-9a62-03de60bf31f7. Create a Python script to perform linear regression to establish how temperature affects dissolved oxygen levels. Test the model on the whole dataset and find the RMSE.
4. Perform logistic regression to classify Cleveland heart disease dataset. Print the feature importance and accuracy. Drop least important attributes one by one and assess how the accuracy and feature importance changes.
5. Find the correlation and covariance between different attributes of Cleveland heart disease dataset. Which are the top 5 attributes closely related to the predicted attribute?
6. Perform Naive Bayes classification on the "glass" dataset from Kaggle. Interpret the performance of the classifier, and evaluate why the accuracy value is what you obtained.
7. Use the "Car Evaluation Dataset" from UCI Machine Learning repository to generate a decision tree and measure the performance.
8. Implement KNN algorithm to classify iris dataset. Print all necessary performance measures.
9. Implement appropriate CNNs to classify (i) MNIST dataset, and (ii) Fashion MNIST dataset. Redesign the CNN with different hyperparameters and evaluate the performance.
10. Implement dimensionality reduction on Car Evaluation dataset from UCI Machine Learning repository using PCA. Try setting number of PCA components from 2 to 5, and identify the composition that gives the best performance among all of them. Find covariance among all features in the original dataset and try to justify the performance.

Reference Books:

1. Aurelien Geron, “Hands–On Machine Learning with Scikit–Learn and TensorFlow”, O’Relily.
2. David dietrich, “EMC education service’s, data science and big data analytics, discovering, analyzing, visualizing, and presenting data”, John Wiley and sons
3. Stuart J. Russell, Peter Norvig, “Artificial Intelligence: A Modern Approach”, Pearson Education.