Pokhara University
Faculty of Science and Technology

Course No.: CMP 360 (2 Credits)	Full marks: 100
Course title: Data Science and Analytics (2-1-2)	Pass marks: 45
Nature of the course: Theory and Practical	Total Lectures: 30 hrs
Level: Bachelor	Program: BE IT / SE / CE/

1. Course Description

This course provides a comprehensive introduction to the field of Data Science and Analytics. Students will learn foundational tools and techniques for collecting, analyzing, and interpreting large datasets, along with practical applications in various domains. The course covers data analysis pipelines, statistical foundations, machine learning techniques, and real-world case studies. By the end of the course, students will be equipped with the skills to apply data science methods to solve practical problems using open-source tools like Python, R, and Weka.

2. General Objectives

- Introduce students to key concepts and tools in data science and analytics.
- Teach students to apply the appropriate data analysis techniques to real-world problems.
- Enable students to understand the assumptions, limitations, and risks of different data analysis methods.
- Provide students with hands-on experience through case studies and project work.

3. Methods of Instruction

Lectures for theoretical foundation.

Tutorial Sessions for interactive learning.

Hands-on **Practical Work** for applied skills.

Project-Based Learning for integrative experience.

Readings and **Assignments** for reinforcement and assessment.

4. Contents in Detail

Objectives	Contents
• Understand the overview of the data analysis process. • Learn the concept of structured and unstructured data • Understand the details of data preprocessing	**Unit I: Data Analysis Pipeline (3 hrs)** 1.1. The Knowledge Discovery from Database Process 1.2. Structured and unstructured data and their examples 1.3. Overview of data preprocessing 1.3.1. Data cleaning 1.3.2. Data integration 1.3.3. Data transformation and discretization
• Learn the descriptive data analysis techniques • Able to infer empirical probability distribution of variables of different types • Implement the knowledge of correlation analysis on variables of different types • Apply the concept of statistical significance and its use in real-world example	**Unit II: Statistical Foundation (8 hrs)** 2.1. Types of Variables: Numeric and Categorical 2.2. Empirical Distribution 2.2.1. Numeric data: histograms, normal, exponential, power laws 2.2.2. Categorical Data: bar plots, binomial distribution, Zipf's law 2.3. Correlation Analysis 2.3.1. Pearson correlation: correlation between numeric variables 2.3.2. Cross-tabulation: correlation between categorical variables 2.3.3. Assessing correlation between numeric and categorical variables 2.4. Statistical Significance: p-value, t-value, chi-squared
• Learn and apply the basic data analysis methods in the context of numerical datasets that are used in the social sciences and business • Implement the knowledge of selecting either among linear or non-linear model for regression in real-world examples • Use PCA as a tool of	**Unit III: Numeric Data (9 hrs)** 3.1.Multivariate Linear Regression 3.1.1. Matrix formulation and OLS estimation 3.1.2. Measures of fit: R-squared and Adjusted R-squared 3.1.3. Multi-collinearity and variance-inflation factors 3.2.Non-parametric regression: Nadaraya-Watson kernel regression 3.2.1. Derivation of the estimator 3.2.2. Rules of setting the appropriate bandwidth size 3.3.Principal Component Analysis (PCA) 3.3.1. Mathematical formulation and relevant

identifying latent variables	derivations 3.3.2. Interpreting the principal components
• Learn to analyze categorical and mixed-type data using machine learning • Assess predictor variable importance in regression and classification models	**Unit IV: Categorical and Mixed-type Data (5)** 4.1. Classification: Decision Trees and the CART algorithm 4.2. Regression: Logistic Regression 4.3. Variable Importance: permutation tests, partial dependence plots 4.4. Clustering: K-means and distance metrics for mixed-type data
• Use ARMA and ARIMA models in time-series modeling problems • Understand the basics of causal analysis	**Unit V: Time and Causality (5)** 5.1.Time series analysis 5.1.1. Autocorrelation and stationarity 5.1.2. ARMA and ARIMA models 5.1.3. Selecting optimal lag length: Akaike and Bayesian Information Criteria 5.2.Causal analysis 5.2.1. Causation vs. correlation 5.2.2. Granger causality: overview only 5.2.3. Causal Directed Acyclic Graph: overview only

5. Practical Works

The laboratory work, consisting of 30 hours per group (with a maximum of 24 students), should focus on applying the key concepts covered in the course to real-world datasets. The labs should emphasize using existing tools and software to identify patterns and relationships in the data, rather than implementing algorithms from scratch. The practical sessions can be organized as a series of lab assignments as following

SN	Description
1	Data Preprocessing: Cleaning and integrating datasets from multiple sources.
2	Plotting and inferring empirical distribution
3	Correlation analysis
4	Applying and interpreting multivariate linear regression
5	Non-parametric regression using kernel methods.
6	Principal Component Analysis (PCA).

7	Decision Trees for classification
8	Logistic-regression
9	ARMA and ARIMA
10	Granger Causal inference

Alternatively, the instructor can organize the practical component as individual student projects. Each project should cover the three main phases of empirical research: data preprocessing, data analysis or data mining, and data visualization. Project topics can either be proposed by the students or assigned by the instructor.

The practical component should resort to open-source languages or tools for data mining, such as R, Python, Octave, or Weka.

6. List of Tutorials

The following tutorial activities of 15 hours per group of maximum 24 students should be conducted to cover the content of this course:

A. Problem solving-based Tutorials: (6 hrs)

a. Calculation of chi-squared statistics and testing variable independence through cross-tabulation.
b. Interpretation of the coefficients and fit-statistics of a linear regression model in a real-world context. The linear regression model should be provided by the instructor.
c. Calculation of the kernel regression estimates on a univariate regression context in a small dataset.
d. Calculation of the principal components from the basic definition on a very small data comprising of three variables.
e. Interpretation of variable importance tests and partial dependence plots for a classification and regression models. The models, test results, and plots should be provided by the instructor.

B. Case Studies: (9 hrs)

a. Case study on the use of basic descriptive and cross-tabulations regarding the Access and use of mass media and ICT in Nepal based on the official report by the National Statistics Office.
b. Case study on the Principal Component Analysis-based estimation of relative wealth from household surveys based on the seminal work of Filmer and Prtichet (2001).
c. Case study on the using logistic regression to predict the occurrence of species in different habitats based on Pearce and Ferrier (2004)

7. Evaluation System and Students' Responsibilities

Evaluation System

The internal evaluation of a student may consist of assignments, attendance, internal assessment, lab reports, project works etc. The internal evaluation scheme for this course is as follows:

Internal Evaluation	**Weight**	**Marks**	**External Evaluation**	**Marks**
Theory		30	Semester-End examination	50
Attendance & Class Participation	10%			
Assignments	20%			
Presentations/Quizzes	10%			
Internal Assessment	60%			
Practical		20		
Attendance & Class Participation	10%			
Lab Report/Project Report	20%			
Practical Exam/Project Work	40%			
Viva	30%			
Total Internal		50		
Full Marks: 50 + 50 = 100				

Student Responsibilities

Each student must secure at least 45% marks separately in internal assessment and practical evaluation with 80% attendance in the class in order to appear in the Semester End Examination. Failing to get such a score will be given NOT QUALIFIED (NQ) to appear for the Semester-End Examinations. Students are advised to attend all the classes, formal exam, test, etc. and complete all the assignments within the specified time period. Students are required to complete all the requirements defined for the completion of the course.

8. Prescribed Books and References

Text Books

1. Johnson, R.A. and Wichern, D.W., 2014. Applied multivariate statistical analysis. PHI Learning Pvt Ltd.
2. Tan, P.N., Steinbach, M. and Kumar, V.,, 2006. Introduction to data mining. Pearson Education, Inc.
3. Han, J., Kamber, M. and Mining, D., 2006. Concepts and techniques. Morgan Kaufmann

References

1. Government of Nepal and UNICEF, 2019. Nepal Multiple Indicator Cluster Survey Report.
2. Filmer, D. and Pritchett, L.H., 2001. Estimating wealth effects without expenditure data—or tears: an application to educational enrollments in states of India. Demography, 38, pp.115-132.
3. Pearce, J. and Ferrier, S., 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecological modelling, 133(3), pp.225-245.
4. Greene, W. H. Econometric Analysis. Fifth Edition. Pearson.