TKR COLLEGE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS)
(Sponsored by TKR Educational Society , Approved by AICTE, Affiliated by JNTUH,
Accredited by NBA & NAAC with 'A' Grade)

B.TECH – COMPUTER SCIENCE & ENGINEERING (DATA SCIENCE)
Course Structure R-20

SEMESTER III

S.No.	Class	Course Code	Name of the Subject	L	T	P	C
1	BS	CBSM5	Statistical Methods	3	0	0	3
2	ES	CESOP1	Introduction to Object-Oriented Programming & Data Structures using Java	3	0	0	3
3	PC	C83PC1	Database Management Systems	3	0	0	3
4	PC	C83PC2	Operating Systems	3	0	0	3
5	PC	C83PC3	R Programming	3	0	0	3
6	PC	C83PC4	Formal Language & Automata Theory	3	0	0	3
7	ES	CESOP2	Introduction to Object-Oriented Programming & Data Structures using Java Lab	0	0	2	1
8	PC	C83PC5	Database Management Systems Lab	0	0	2	1
9	PC	C83PC6	R Programming Lab	0	0	2	1
10	MC	MC003	Cultural Activity	0	0	0	Satisfactory
Total Credits				**18**	**0**	**6**	**21**

Mandatory Course: Cultural Activity

The student should participate in culture activity (Music/Dance/Singing/etc.) conducted by the College, student should produce the participation certificate for clearing this course.

SEMESTER IV

S.No.	Class	Course Code	Name of the Subject	L	T	P	C
1	BS	CBSM6	Probability & Algebra	3	0	0	3
2	PC	C84PC1	Data Warehousing & Data Mining	3	0	0	3
3	PC	C84PC2	Information Security	3	0	0	3
4	PC	C84PC3	Design and Analysis of Algorithms	3	0	0	3
5	PC	C84PC4	Data Visualization	3	0	0	3
6	PC	C84PC5	Python Programming	3	0	0	3
7	PC	C84PC6	Data Warehousing & Data Mining Lab	0	0	2	1
8	PC	C84PC7	Data Visualization Lab	0	0	2	1
9	PC	C84PC8	Python Programming Lab	0	0	2	1
10	MC	MC004	Video with Social Messages	0	0	0	Satisfactory
Total Credits				**18**	**0**	**6**	**21**

Mandatory Course: Video with Social Messages

Student should make video with social messages. This has to be uploaded in the youtube.com, by maintaining the terms and conditions of youtube.com. Student should produce youtube.com link with screen shot for clearing this mandatory course.

TKR COLLEGE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS)
(Sponsored by TKR Educational Society , Approved by AICTE, Affiliated by JNTUH,
Accredited by NBA & NAAC with 'A' Grade)

CSE (DATA SCIENCE)

B.Tech IV Semester **L/T/P/C**
3 /0/ 0/ 3

PROBABILITY & ALGEBRA (CBSM6)

Course Objectives:

To learn:

1. Concepts of Basic probability.
2. Random variables that describe randomness or an uncertainty in certain realistic situation.
3. The study of discrete and continuous distributions predominantly describes important probability distributions.
4. To relate practical examples to the appropriate set, function or relation model, and interpret the associated operations and terminology in context.
5. Introduce the concepts of semi groups, monoids, groups, sub-groups, abelian groups, Isomorphism and homomorphism of groups.

Course Outcomes:

After learning the contents of this course, the student will be able to
CO1: Learn the concept of basic probability to solve the real life problems.
CO2: To solve problems on discrete and continuous random variables.
CO3: Learn various discrete and continuous probability distribution and their properties.
CO4: Solve problems based on area properties of standard normal distribution.
CO5: Illustrate the basic terminology of functions, relations, sets, and demonstrate knowledge of their associated operations.
CO6: Understand the importance of algebraic properties with regard to working within various number systems.

UNIT I
Probability

Basic concepts of probability, Axiomatic definition of probability, Addition theorem, conditional probability, multiplication theorem, Independent events, Baye's theorem.

UNIT II
Random Variables

Random variables –discrete and continuous, Mathematical expectation, Variance, co-variance, joint and marginal probability density function, statistical independence.

PROBABILITY & ALGEBRA (CBSM6)

UNIT III
Distributions

Probability mass function, density function of Binomial, Poisson and Normal distributions related properties.

UNIT IV
Relations

Properties of Binary relations, equivalence, transitive closure, compatibility and partial ordering Relations, Hasse diagram.
Functions: Inverse function, composition of functions, recursive functions.

UNIT V
Groups

Algebraic structures, examples and general properties, Semi groups and monoids, Groups, Sub groups, cosets and Lagranges theorem, homomorphism, and isomorphism of groups, cyclic groups, permutation groups.

Text Books:

1. Probability & Statistics for Engineers by G.S.S. Bhisma Rao, SciTech Publications.
2. Discrete Mathematics for Computer scientists & Mathematicians, J. L. Mott, A. Kandel, T.P.Baker.

Reference Books:

1. W.Feller- An introduction to probability theory and its applications- Vol.1- 3rd edition Wiley-1968.
2. Probability & Statistics for Engineers, Millers and John E. Freund, Prentice Hall of India.
3. Discrete mathematical structures theory and applications- malik & Sen Cengage.

TKR COLLEGE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS)
(Sponsored by TKR Educational Society , Approved by AICTE, Affiliated by JNTUH,
Accredited by NBA & NAAC with 'A' Grade)

CSE (DATA SCIENCE)

B.Tech IV Semester **L/T/P/C**
3 /0/ 0/ 3

DATA WAREHOUSING & DATA MINING (C84PC1)

Course Objectives:

Study data warehouse principles and its working learn data mining concepts understand association rules mining. Discuss classification algorithms learn how data is grouped using clustering techniques.

Course Outcomes:

1. Be familiar with mathematical foundations of data mining tools.
2. Understand and implement classical models and algorithms in data warehouses and data mining.
3. Characterize the kinds of patterns that can be discovered by association rule mining, classification and clustering
4. Master data mining techniques in various applications like social, scientific and environmental context
5. Develop skill in selecting the appropriate data mining algorithm for solving practical problems.

UNIT I
Data Warehouse

Introduction to Data warehouse, Difference between operational database systems and data warehouses. Data warehouse Characteristics, Data warehouse Architecture and its Components, Extraction – Transformation – Loading, Logical (Multi – Dimensional), Data Modelling, Schema Design, Star, Snow Flake Schema and Fact Constellation, Fact Table, Fully Addictive, Semi – Addictive, Non Addictive Measures; Fact Constellation, Fact Table, Fully Addictive, Semi – Addictive, Non Addictive Measures; Fact – Less – Facts, Dimension Table Characteristics; OLAP Cube, OLAP Operations, OLAP Server Architecture – ROLAP, MOLAP and HOLAP

UNIT II
Introduction to Data Mining

Introduction, What is Data Mining, Definition, KDD, Challenges, Data Mining Tasks, Data Preprocessing, Data Cleaning, Missing data, Dimensionality Reduction, Feature Subset Selection, Discretization and Binarization, Data Transformation; Measures of Similarity and Dissimilarity- Basics.

Unit III
Association Rules

Problem Definition, Frequent Item Set Generation, The APRIORI Principle, Support and Confidence Measures, Association Rule Generation; APRIOIRI Algorithm, The Partition Algorithms, FP-Growth Algorithms, Compact Representation of Frequent Item Set, Maximal Frequent Item Set, Closed Frequent Item Set.

DATA WAREHOUSING & DATA MINING (C84PC1)

UNIT IV
Classification

Problem Definition, General Approaches to solving a classification problem, Evaluation of Classifiers, Classification techniques, Decision Trees-Decision tree Construction, Methods for Expressing attribute test conditions, Measures for Selecting the Best Split, Algorithm for Decision tree Induction; Naïve – Bayes Classifier, Bayesian Belief Networks; K- Nearest neighbor classification-Algorithm and Characteristics.

UNIT V
Clustering

Problem Definition, Clustering Overview, Evaluation of Clustering Algorithms, Partitioning Clustering-K-Means Algorithm, K-Means Additional issues, PAM Algorithm; Hierarchical Clustering-Agglomerative Methods and divisive methods, Basic Agglomerative Hierarchical Clustering Algorithm, Specific techniques, Key Issues in Hierarchical Clustering, Strengths and Weakness; Outlier Detection.

Text Books:

1. Data Mining- Concepts and Techniques- Jiawei Han, Micheline Kamber, Morgan Kaufmann Publishers, Elsevier 2 Edition, 2006.
2. Introduction to Data Mining, Pang-Ning Tan, Vipin Kumar, Michael Steinbanch, Pearson Education.

Reference Books:

1. Data Mining Techniques, Arun K Pujari, 3rd Edition, Universities Press.
2. Data Warehousing Fundamentals, Paulraj Ponnaiah, Wiley Student Edition.
3. The Data Warehousing Life Cycle Toolkit – Ralph Kimbal. Wiley Student Edition.
4. Data Mining, Vikaram Pudi, P Radha Krishna, Oxford University Press.

TKR COLLEGE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS)
(Sponsored by TKR Educational Society , Approved by AICTE, Affiliated by JNTUH,
Accredited by NBA & NAAC with 'A' Grade)

CSE (DATA SCIENCE)

B.Tech IV Semester **L/T/P/C**
3 /0/ 0/ 3

INFORMATION SECURITY (C84PC2)

Course Objective:

To understand and learn the objectives of Network security, Cryptographic algorithms.

Course Outcomes:

After completion of the course student will be able to

1. Understand the security concepts and classical encryption techniques.
2. Understand the symmetric and asymmetric key algorithms
3. Understand the authentication and hash algorithms.
4. Understand cryptographic algorithms for web, E-mail and security issues.
5. Understand the IP security and system security.

UNIT I
Security Concepts

Introduction, security trends, OSI Architecture, security attacks, security services, security mechanisms, A Model for Network Security.
Cryptography Concepts and Techniques: Introduction, Plain Text and cipher text, substitution techniques(Caesar cipher, Playfair cipher, Hill cipher), transposition techniques, steganography.

UNIT II
Symmetric Key Ciphers

Block Cipher principles, DES, AES, Block Cipher Modes of Operation, Stream ciphers, RC4.
Asymmetric Key Ciphers: Principles of public key cryptosystems, RSA algorithm, Difffie-Hellman Key Exchange, Elliptic Curve Cryptography and Arithmetic.

UNIT III
Cryptographic Hash Functions

Authentication requirements and Functions, Message Authentication Code, Secure Hash Algorithm (SHA-512), Message authentication codes: HMAC, CMAC, Digital signatures,
AUTHENTICATION APPLICATIONS: Kerberos, X.509 Authentication Service, Public – Key Infrastructure.

INFORMATION SECURITY (C84PC2)

UNIT IV
Web Security

Web security considerations, Secure Socket Layer, and Transport Layer Security, Secure Electronic Transaction.
E-Mail Security: Pretty Good Privacy, S/MIME.

UNIT V
IP Security

IP Security overview, IP Security architecture, Authentication Header, Encapsulating Security Payload, Combining Security Associations, Key Management.
System Security: Intruders, Intrusion Detection, Password Management.

Text Books:

1. Cryptography and Network Security - Principles and Practice: William Stallings, Pearson Education, 4th Edition
2. Cryptography and Network Security: Atul Kahate, Mc Graw Hill, 3rd Edition

Reference Books:

1. Cryptography and Network Security: C K Shyamala, N Harini, Dr T R Padmanabhan, Wiley India, 1st Edition.
2. Cryptography and Network Security: Forouzan, Mukhopadhyay, Mc Graw Hill,3rd Edition.
3. Information Security, Principles, and Practice: Mark Stamp, Wiley India.
4. Principles of Computer Security: WM, Arthur Conklin, Greg White, TMH.
5. Introduction to Network Security: Neal Krawetz, CENGAGE Learning.
6. Network Security and Cryptography: Bernard Menezes, CENGAGE Learning.

TKR COLLEGE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS)
(Sponsored by TKR Educational Society , Approved by AICTE, Affiliated by JNTUH,
Accredited by NBA & NAAC with 'A' Grade)

CSE (DATA SCIENCE)

B.Tech IV Semester **L/T/P/C**
3 /0/ 0/ 3

DESIGN AND ANALYSIS OF ALGORITHMS (C84PC3)

Course Objective:

To understand the design paradigms for developing an algorithm and analyzing it for a given problem.

Course Outcomes:

1. Argue the correctness of algorithms using inductive proofs and invariants.
2. Apply important algorithmic design paradigms and methods of analysis.
3. Synthesize efficient algorithms in common engineering design situations such as the greedy, divide and conquer, dynamic programming, backtracking and branch-bound.
4. Explain the different ways to analyze randomized algorithms (expected running time, probability of error)
5. Differentiate between tractable and intractable problems.

UNIT I
Introduction

Algorithm definition, Algorithm Specification, Performance Analysis-Space complexity, Time complexity, Randomized Algorithms. Divide and conquer- General method, applications – Binary search, Merge sort, Quick sort, Strassen's Matrix Multiplication.

UNIT II
Disjoint Set Operations

Disjoint set operations, union and find algorithms, AND/OR graphs, Connected Components and Spanning trees, Bi-connected components Backtracking-General method, applications the 8-queen problem, sum of subsets problem, graph coloring, Hamiltonian cycles.

UNIT III
Greedy Method

General method, applications- Knapsack problem, Job sequencing with deadlines, Minimum cost spanning trees, Single source shortest path problem.

UNIT IV
Dynamic Programming

General Method, applications- Chained matrix multiplication, All pairs shortest path problem, Optimal binary search trees, 0/1 knapsack problem, Reliability design, Travelling sales person problem.

DESIGN AND ANALYSIS OF ALGORITHMS (C84PC3)

UNIT V
Branch and Bound

General Method, applications-0/1 Knapsack problem, LC Branch and Bound solution, FIFO Branch and Bound solution, Traveling sales person problem. NP-Hard and NP Complete problems- Basic concepts, Non-deterministic algorithms, NP – Hard and NP-Complete classes, Cook's theorem.

Text Books:

1. Fundamentals of Computer Algorithms, 2nd Edition, Ellis Horowitz, Sartaj Sahni and S.Rajasekharan, Universities Press.
2. Design and Analysis of Algorithms, P. H. Dave, H. B. Dave, 2nd edition, Pearson Education.

Reference Books:

1. Algorithm Design: Foundations, Analysis and Internet examples, M. T. Goodrich and R. Tomassia, John Wiley and sons.
2. Design and Analysis of Algorithms, S. Sridhar, Oxford Univ. Press
3. Design and Analysis of algorithms, Aho, Ullman and Hopcroft, Pearson Education.
4. Foundations of Algorithms, R. Neapolitan and K. Naimipour, 4th edition, Jones and Bartlett Student edition.
5. Introduction to Algorithms, 3rd Edition, T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein PHI.

TKR COLLEGE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS)
(Sponsored by TKR Educational Society , Approved by AICTE, Affiliated by JNTUH,
Accredited by NBA & NAAC with 'A' Grade)

CSE (DATA SCIENCE)

B.Tech IV Semester **L/T/P/C**
3 /0/ 0/ 3

DATA VISUALIZATION (C84PC4)

Course Objectives:

To understand the visual representation of structured and un structured data.

Course Outcomes:

After completion of course, the students will be able to

1. Understand the visualization and Data basics
2. Understand the Visualization process and know the representation of Spatial & Geo spatial data
3. Analyze various Visualization techniques for Multivariate data and other structures of data
4. Interacting the different operators and different data spaces
5. Design effective visualization of modern toolkits

UNIT I
Introduction

What is Visualization, History, Relationship visualization with other fields, The visualization Process, Pseudocode Conventions, The Scatter plot
Data Foundations: Types of Data, Structure within and between the records, Data Processing.

UNIT II
Visualization Foundations

The Visual Process, Semiology of Graphical Symbols, The Eight Visual Variables, Historical Perspective, Taxonomies.
Visualization Techniques for Spatial Data: One-Dimensional Data, Two-Dimensional Data, Three-Dimensional Data, Dynamic Data, Combining Techniques.
Visualization Techniques for Geospatial Data: Visualizing Spatial Data, Visualization of Point Data, Visualization of Line Data, Visualization of Area Data.

UNIT III
Visualization Techniques for Multivariate Data

Point-Based Techniques, Line-Based Techniques, Region-Based Techniques, Combinations of Techniques, Visualization Techniques for Trees, Graphs, and Networks: Displaying Hierarchical Structures, Displaying Arbitrary Graphs/Networks.

DATA VISUALIZATION (C84PC4)

UNIT IV
Text and Document Visualization

Levels of Text Representation, The Vector Space Model, Single Document Visualizations, Document Collection Visualizations.
Interaction Concepts: Interaction Operators, Interaction Operands and Spaces, A Unified Framework.
Interaction Techniques: Object Space, Data Space, Attribute Space, Data Structure Space, Visualization Structure Space, Animating Transformations, Interaction Control.

UNIT V
Designing Effective Visualizations

Steps in Designing Visualizations, Problems in Designing Effective Visualizations
Comparing and Evaluating Visualization Techniques: User Tasks, User Characteristics, Data Characteristics, Visualization Characteristics, Structures for Evaluating Visualizations.
Visualization Systems: Systems Based on Data Type, Systems Based on Analysis Type, Text Analysis and Visualization, Modern Integrated Visualization Systems, Toolkits

Text Books:

1. Interactive Data Visualization Foundations, Techniques, and Applications by Grinstein, Georges Keim, Daniel Ward, Matthew O , CRC Press Taylor & Francis Group.
2. Digital Image. Processing. Third Edition. Rafael C. Gonzalez. University of Tennessee. Richard E. Woods. NledData Interactive. Pearson International Edition.

TKR COLLEGE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS)
(Sponsored by TKR Educational Society , Approved by AICTE, Affiliated by JNTUH,
Accredited by NBA & NAAC with 'A' Grade)

CSE (DATA SCIENCE)

B.Tech IV Semester **L/T/P/C**
3 /0/ 0/ 3

PYTHON PROGRAMMING (C84PC5)

Course Objective:

Enable the student to do Python Programming which includes Regular Expressions and GUI

Course Outcomes:

After completion of course the student will be able to

1. Examine Python syntax and semantics and be fluent in the use of Python flow control and functions.
2. Demonstrate proficiency in handling Strings and File Systems.
3. Create, run and manipulate Python Programs using core data structures like Lists, Dictionaries and use Regular Expressions.
4. Interpret the concepts of Object-Oriented Programming as used in Python.
5. Implement exemplary applications related to Network Programming, Web Services and Databases in Python.

UNIT I
Introduction

Introduction to Python, History, Need of Python Programming, features Applications, python environment setup, Basic syntax, Variables, Data Types, Keywords, Input-Output, Indentation, script structure, Running Python Scripts.
Operators: Arithmetic Operators, Comparison (Relational) Operators, Assignment Operators, Logical Operators, Bitwise Operators, Membership Operators, Identity Operators, Expressions and order of evaluations, Conditional statements if, if-else Looping Control Structures for, while Control Statements: Break, Continue, Pass.

UNIT II
Functions

Defining Functions, Calling Functions, Passing Arguments, Keyword Arguments, Default Arguments, Variable-length arguments, Anonymous Functions, Fruitful Functions (Function Returning Values), Scope of the Variables in a Function - Global and Local Variables.
Data Structures : Lists, Tuples, dictionaries, sets, Sequences, Comprehensions.

PYTHON PROGRAMMING (C84PC5)

UNIT III
Regular Expressions

Introduction/Motivation , Special Symbols and Characters, REs and Python.
OBJECT ORIENTED PROGRAMMING IN PYTHON
Classes, 'self-variable', Methods, Constructor Method, Inheritance, Overriding Methods, Data hiding. ERROR AND EXCEPTIONS Difference between an error and Exception, Handling exceptions, try, except block, Raising Exceptions and User Defined Exceptions.

UNIT IV
Files

File input/output, Text processing file functions.
MODULES and Introduction to Packages, Creating modules, import statement, from. Name spacing, Packages, using packages, implementing packages: numpy, iterator tools, scipy, matplotlib.

UNIT V
GUI Programming

Introduction, Tkinter and Python Programming, Brief Tour of other GUIs, Related Modules and other GUIs.
Database Programming: Introduction, Python Database, Application Programmer's Interface (DB-API), Object Relational Managers (ORMs), Related Modules.

Text Book:

1. Core Python Programming, Wesley J. Chun, Second Edition, Pearson.

Reference Books:

1. Allen Downey, "Think Python", Second Edition , Green Tea Press.
2. Introduction to Computation & Programming Using Python, Spring 2013 Edition, By John V.Guttag.
3. 3. Programming in Python 3: A Complete Introduction to the Python Language (Developer's Library), by Mark Summerfield, 2nd Edition.

TKR COLLEGE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS)
(Sponsored by TKR Educational Society , Approved by AICTE, Affiliated by JNTUH,
Accredited by NBA & NAAC with 'A' Grade)

CSE (DATA SCIENCE)

B.Tech IV Semester **L/T/P/C**
0 /0/ 2/ 1

DATA WAREHOUSING AND DATA MINING LAB (C84PC6)

Course Objective:

Learn how to build a data warehouse and query it, perform data mining tasks using a data mining toolkit and understand the data sets and data preprocessing.

Course Outcomes:

After completion of the course, the student will be able to

1. Ability to understand the various kinds of tools
2. Demonstrate the classification, clustering and etc. in large data sets.
3. Ability to add mining algorithms as a component to the exiting tools.
4. Ability to apply mining techniques for realistic data

List of experiments and Tasks

Experiment-1: Build Data Warehouse and Explore WEKA

A. Build Data Warehouse/Data Mart (using open source tools like Pentaho Data Integration Tool, Pentaho Business Analytics; or other data warehouse tools like Microsoft-SSIS, Informatica, Business Objects, etc.,)

(i) Identify source tables and populate sample data.

(ii) Design multi-dimensional data models namely Star, Snowflake and Fact Constellation schemas for any one enterprise (ex. Banking, Insurance, Finance, Healthcare, manufacturing, Automobiles, sales etc).

(iii) Write ETL scripts and implement using data warehouse tools

(iv) Perform Various OLAP operations such as slice, dice, roll up, drill up and pivot.

(v). Explore visualization features of the tool for analysis like identifying trends etc.

B. Explore WEKA Data Mining/Machine Learning Toolkit.

(i) Downloading and/or installation of WEKA data mining toolkit.

(ii) Understand the features of WEKA tool kit such as Explorer, Knowledge flow interface, Experimenter, command-line interface.

(iii) Navigate the options available in the WEKA(ex.select attributes panel, preprocess panel, classify panel, cluster panel, associate panel and visualize)

(iv) Study the ARFF file format

(v) Explore the available data sets in WEKA.

(vi) Load a data set (ex.Weather dataset,Iris dataset,etc.)

(vii) Load each dataset and observe the following:

(vii.i) List attribute names and they types

(vii.ii) Number of records in each dataset.

(vii.iii) Identify the class attribute (if any)

(vii.iv) Plot Histogram

(vii.v) Determine the number of records for each class

(vii.vi) Visualize the data in various dimensions

DATA WAREHOUSING AND DATA MINING LAB (C84PC6)

Experiment-2: Perform data preprocessing tasks and Demonstrate performing association rule mining on data sets

A. Explore various options in Weka for preprocessing data and apply (like Discretization Filters, Resample filter, etc.) in each dataset.
B. Load each dataset into Weka and run Apriori algorithm with different support and confidence values. Study the rules generated.
C. Apply different discretization filters on numerical attributes and run the Apriori association rule algorithm. Study the rules generated. Derive interesting insights and observe the effect of discretization in the rule generation process.

Experiment-3: Demonstrate performing classification on data sets.

A. Load each dataset into Weka and run id3, j48 classification algorithm, study the classifier output. Compute entropy values, Kappa statistic.
B. Extract if-then rues from decision tree generated by classifier, Observe the confusion matrix and derive Accuracy, F- measure, TPrate, FPrate , Precision and recall values. Apply cross-validation strategy with various fold levels and compare the accuracy results.
C. Load each dataset into Weka and perform Naïve-bayes classification and k-Nearest Neighbor classification, Interpret the results obtained.
D. Plot RoC Curves.
E. Compare classification results of ID3,J48, Naïve-Bayes and k-NN classifiers for each dataset , and reduce which classifier is performing best and poor for each dataset and justify.

Experiment-4: Demonstrate Performing Clustering on Data Sets Clustering Tab

A. Load each dataset into Weka and run simple k-means clustering algorithm with different values of k(number of desired clusters). Study the clusters formed. Observe the sum of squared errors and centroids, and derive insights.
B. Explore other clustering techniques available in Weka.
C. Explore visualization features of weka to visualize the clusters. Derive interesting insights and explain.

Experiment-5: Demonstrate Performing Regression on Data Sets

A. Load each dataset into Weka and build Linear Regression model. Study the cluster formed. Use training set option. Interpret the regression model and derive patterns and conclusions from the regression results.
B. Use options cross-validation and percentage split and repeat running the Linear Regression Model. Observe the results and derive meaningful results.
C. Explore Simple linear regression techniques that only looks at one variable.

Experiment-5: Sample Programs using German Credit Data.

Task 1: Credit Risk Assessment

Description: The business of banks is making loans. Assessing the credit worthiness of an applicant is of crucial importance. You have to develop a system to help a loan officer decide whether the credit of a customer is good or bad. A bank's business rules regarding loans must consider two opposing factors. On the one hand, a bank wants to make as many loans as possible.

Interest on these loans is the banks profit source. On the other hand, a bank cannot afford to make too many bad loans. Too many bad loans could lead to the collapse of the bank. The bank's loan policy must involved a compromise. Not too strict and not too lenient. To do the assignment, student first and foremost need some knowledge about the world of credit. Student can acquire such knowledge in a number of ways.

DATA WAREHOUSING AND DATA MINING LAB (C84PC6)

1. Knowledge engineering: Find a loan officer who is willing to talk. Interview him/her and try to represent him/her knowledge in a number of ways.
2. Books: Find some training manuals for loan officers or perhaps a suitable textbook on finance. Translate this knowledge from text form to production rule form.
3. Common sense: Imagine yourself as a loan officer and make up reasonable rules which can be used to judge the credit worthiness of a loan applicant.
4. Case histories: Find records of actual cases where competent loan officers correctly judged when and not to. Approve a loan application.

The German Credit Data

Actual historical credit data is not always easy to come by because of confidentiality rules. Here is one such data set. Consisting of 1000 actual cases collected in Germany.

In spite of the fact that the data is German, you should probably make use of it for this assignment(Unless you really can consult a real loan officer!)

There are 20 attributes used in judging a loan applicant (ie., 7 Numerical attributes and 13 Categorical or Nominal attributes). The goal is to classify the applicant into one of two categories. Good or Bad.

Subtasks:

1. List all the categorical (or nominal) attributes and the real valued attributes separately.

2. What attributes do you think might be crucial in making the credit assessment? Come up with some simple rules in plain English using your selected attributes.

3. One type of model that you can create is a Decision tree . train a Decision tree using the complete data set as the training data. Report the model obtained after training.

4. Suppose you use your above model trained on the complete dataset, and classify credit good/bad for each of the examples in the dataset. What % of examples can you classify correctly?(This is also called testing on the training set) why do you think can not get 100% training accuracy?

5. Is testing on the training set as you did above a good idea? Why or why not?

6. One approach for solving the problem encountered in the previous question is using cross-validation? Describe what is cross validation briefly. Train a decision tree again using cross validation and report your results. Does accuracy increase/decrease? Why?

7. Check to see if the data shows a bias against “foreign workers” or “personal-status”. One way to do this is to remove these attributes from the data set and see if the decision tree created in those cases is significantly different from the full dataset case which you have already done. Did removing these attributes have any significantly effect? Discuss.

8. Another question might be, do you really need to input so many attributes to get good results? May be only a few would do. For example, you could try just having attributes 2,3,5,7,10,17 and 21. Try out some combinations.(You had removed two attributes in problem 7. Remember to reload the arff data file to get all the attributes initially before you start selecting the ones you want.)

DATA WAREHOUSING AND DATA MINING LAB (C84PC6)

9. Sometimes, The cost of rejecting an applicant who actually has good credit might be higher than accepting an applicant who has bad credit. Instead of counting the misclassification equally in both cases, give a higher cost to the first case (say cost 5) and lower cost to the second case. By using a cost matrix in weak. Train your decision tree and report the Decision Tree and cross validation results. Are they significantly different from results obtained in problem 6.

10. Do you think, it is a good idea to predict simple decision trees instead of having long complex decision tress? How does the complexity of a Decision Tree relate to the bias of the model?

11. You can make your Decision Trees simpler by pruning the nodes. One approach is to use Reduced Error Pruning. Explain this idea briefly. Try reduced error pruning for training your Decision Trees using cross validation and report the Decision Trees you obtain? Also Report your accuracy using the pruned model Does your Accuracy increase?

12. How can you convert a Decision Tree into "if-then-else rules". Make up your own small Decision Tree consisting 2-3 levels and convert into a set of rules. There also exist different classifiers that output the model in the form of rules. One such classifier in weka is rules, train this model and report the set of rules obtained. Sometimes just one attribute can be good enough in making the decision, yes, just one! Can you predict what attribute that might be in this data set? OneR classifier uses a single attribute to make decisions (it chooses the attribute based on minimum error).Report the rule obtained by training a oneR classifier. Rank the performance of j48, OneR.

TKR COLLEGE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS)
(Sponsored by TKR Educational Society , Approved by AICTE, Affiliated by JNTUH,
Accredited by NBA & NAAC with 'A' Grade)

CSE (DATA SCIENCE)

B.Tech IV Semester **L/T/P/C**
0 /0/ 2/ 1

DATA VISUALIZATION LAB (C84PC7)

Course Objectives:

To obtain practical experience using Tableau public or similar tools.

Course Outcomes:

After completion of the lab student will be able to:
Visualize the different types of data.

List of Sample Problems:

1. Tableau/ Power BI or similar tools setup for Data Visualization (Importing packages etc.…).
2. Extracting and operations of data from different sources.
3. Working on worksheets.
4. Applying the Calculations (Operators, Functions, Numerical Calculations, String, Date, Table).
5. Usage of Different types of Filter & Sort.
6. Construction of Charts (Line, BAR, etc.).
7. Creation of Dashboard (Optional).

TKR COLLEGE OF ENGINEERING AND TECHNOLOGY
(AUTONOMOUS)
(Sponsored by TKR Educational Society , Approved by AICTE, Affiliated by JNTUH,
Accredited by NBA & NAAC with 'A' Grade)

CSE (DATA SCIENCE)

B.Tech IV Semester **L/T/P/C**
0 /0/ 2/ 1

PYTHON PROGRAMMING LAB (C84PC8)

Course Objectives:

1. To be able to introduce core programming basics and program design with functions using Python programming language.
2. To understand a range of Object-Oriented Programming, as well as in-depth data and information processing techniques.
3. To understand the high-performance programs designed to strengthen the practical expertise.

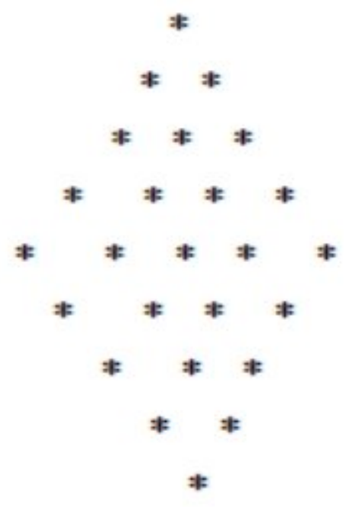
Course Outcomes:

1. Student should be able to understand the basic concepts scripting and the contributions of scripting language.
2. Ability to explore python especially the object-oriented concepts, and the built-in objects of Python.
3. Ability to create practical and contemporary applications such as TCP/IP network programming, Web applications, discrete event simulations.

List of Programs:

1. Write a program to demonstrate different number data types in Python.
2. Write a program to perform different Arithmetic Operations on numbers in Python.
3. Write a program to create, concatenate and print a string and accessing sub-string from a given string.
4. Write a python script to print the current date in the following format "Sun May 29 02:26:23 IST 2017"
5. Write a program to create, append, and remove lists in python.
6. Write a program to demonstrate working with tuples in python.
7. Write a program to demonstrate working with dictionaries in python.
8. Write a python program to find largest of three numbers.
9. Write a Python program to convert temperatures to and from Celsius, Fahrenheit. [Formula: c/5 = f-32/9]
10. Write a Python program to construct the following pattern, using a nested for loop

```
*
* *
* * *
* * * *
* * * * * *
* * * * *
  * * *
    * *
      *
```



PYTHON PROGRAMMING LAB (C84PC8)

11. Write a Python script that prints prime numbers less than 20.
12. Write a python program to find factorial of a number using Recursion.
13. Write a program that accepts the lengths of three sides of a triangle as inputs. The program output should indicate whether or not the triangle is a right angled triangle (Recall from the Pythagorean Theorem that in a right angled triangle, the square of one side equals the sum of the squares of the other two sides).
14. Write a python program to define a module to find Fibonacci Numbers and import the module to another program.
15. Write a python program to define a module and import a specific function in that module to another program.
16. Write a script named copyfile.py. This script should prompt the user for the names of two text files. The contents of the first file should be input and written to the second file.
17. Write a program that inputs a text file. The program should print all of the unique words in the file in alphabetical order.
18. Write a Python class to convert an integer to a roman numeral.
19. Write a Python class to implement pow(x, n).
20. Write a Python class to reverse a string word by word.
21. Write a python program to demonstrate GUI form.
22. Write a python program to Create a database and perform SQL commands.