COURSE INSTRUCTOR NAME: ACADEMIC YEAR:2023-24

SUBJECT NAME:Data Analytics Using R

EMAIL-ID: CLASS ROOM NO:D205

CONTACT NO: 99410 07874

SEM START DATE AND END DATE: 21-8-23 T0 23-12-23

CONTENTS OF COURSE FILE

1. **Department vision & mission**
2. **List of PEOs, POs, PSOs**
3. **List of Cos (Action verbs as per blooms with BTL)**
4. **Syllabus copy and suggested or reference books**
5. **Individual Time Table**
6. **Session plan/ lesson plan**
7. **Session execution log**
8. **Lecture notes(handwritten or softcopy printout-5 units)**
9. **Assignment Questions with (original or Xerox of mid 1 and mid 2 assignment samples)**
10. **Mid exam question papers with (Xerox of mid 1 and mid 2 script samples)**
11. **Scheme of evaluation**
12. **Mapping of Cos with Pos and PSOs**
13. **COs, POs, PSOs Justification**
14. **Attainment of Cos, Pos and PSOs (Excel sheet)**
15. **Previous year question papers**
16. **Power point presentations (PPTs)**
17. **Innovative Teaching method**
18. **References (Textbook/Websites/Journals)**

HOD

DEPARTMENT VISION & MISSION

- **VISION**

To produce globally competent and industry-ready graduates in Computer Science & Engineering by imparting quality education with the know-how of cutting-edge technology and holistic personality.

- **MISSION**

1. To offer high-quality education in Computer Science & Engineering in order tobuild core competence for the graduates by laying a solid foundation in Applied Mathematics and program frameworkwith a focus on concept building.
2. The department promotes excellence in teaching, research, and collaborative activities to prepare graduates for a professional career or higher studies.
3. Creating an intellectual environment for developing logical skills and problem-solving strategies, thus developing, an able and proficient computer engineer to compete in the current global scenario.

2. LIST OF PEOS, POS & PSOs

2.1 PROGRAM EDUCATIONAL OBJECTIVES (PEO):

PEO 1: Excel in professional career and higher education by acquiring knowledge of mathematical computing and engineering principles.
PEO 2: To provide an intellectual environment for analyzing and designing computing systems for technical needs.
PEO 3: Exhibit professionalism to adapt current trends using lifelong learning with legal and ethical responsibilities.
PEO 4: To produce responsible graduates with effective communication skills and multidisciplinary practices to serve society and preserve the environment.

2.2. PROGRAM OUTCOMES:

- **PO1. Engineering Knowledge:**

  An ability to apply knowledge of computing, mathematics, science and engineering fundamentals appropriate to the discipline.

- **PO2.Problem Analysis:**

  An ability to analyze a problem, and identify and formulate the computing requirements appropriate to its solution.

- **PO3.Design/Development Of Solutions:**

  An ability to design, implement, and evaluate a computer-based system, process, component, or program to meet desired needs with appropriate consideration for public health and safety, cultural, societal and environmental considerations.

- **PO4**. **Conduct Investigations Of Computer Programs:**

  An ability to design and conduct experiments, as well as to analyze and interpret data.

- **PO5**. **Modern Tool Usage:**

  An ability to use current techniques, skills, and modern tools necessary for computing practice.

- **PO6.The Engineer And Society:**

  An ability to analyze the local and global impact of computing on individuals, organizations, and society.

- **PO7. Environment And Sustainability:**

  Knowledge of contemporary issues.

- **PO8. Ethics**:

  An understanding of professional, ethical, legal, security and social issues and responsibilities.

- **PO9. Individual And Team Work;**

  An ability to function effectively individually and on teams, including diverse and multidisciplinary, to accomplish a common goal.

- **PO10. Communication**:

  An ability to communicate effectively with a range of audiences.

- **PO11. Project Management And Finance** :

  An understanding of engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects.

- **PO12.life long-learning**:

  Recognition of the need for and an ability to engage in continuing professional development.

2.3. PROGRAM SPECIFIC OUTCOMES (PSOs)

PSO1: **Professional Skills and Foundations of Software development:** Ability to analyze, design and develop applications by adopting the dynamic nature of Software developments.

PSO2: **Applications of Computing and Research Ability**: Ability to use knowledge in cutting edge technologies in identifying research gaps and to render solutions with innovative ideas.

3. LIST OF COs COURSE OUTCOMES:

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

CO1	**Understanding** the building blocks of R-Programming **(Understanding)**
CO2	**Apply** critical R-Programming concepts to handle the data **(Applying)**
CO3	**Apply** statistical concepts on real data **(Applying)**
CO4	**Analyze** logistic and linear regression models on real data **(Analyze**)
CO5	**Create** decision trees to classify the data **(Create)**

REVISED Bloom's Taxonomy Action Verbs

Definitions	I. Remembering	II. Understanding	III. Applying	IV. Analyzing	V. Evaluating	VI. Creating
Bloom's Definition	Exhibit memory of previously learned material by recalling facts, terms, basic concepts, and answers.	Demonstrate understanding of facts and ideas by organizing, comparing, translating, interpreting, giving descriptions, and stating main ideas.	Solve problems to new situations by applying acquired knowledge, facts, techniques and rules in a different way.	Examine and break information into parts by identifying motives or causes. Make inferences and find evidence to support generalizations.	Present and defend opinions by making judgments about information, validity of ideas, or quality of work based on a set of criteria.	Compile information together in a different way by combining elements in a new pattern or proposing alternative solutions.
Verbs	• Choose • Define • Find • How • Label • List • Match • Name • Omit • Recall • Relate • Select • Show • Spell • Tell • What • When • Where • Which • Who • Why	• Classify • Compare • Contrast • Demonstrate • Explain • Extend • Illustrate • Infer • Interpret • Outline • Relate • Rephrase • Show • Summarize • Translate	• Apply • Build • Choose • Construct • Develop • Experiment with • Identify • Interview • Make use of • Model • Organize • Plan • Select • Solve • Utilize	• Analyze • Assume • Categorize • Classify • Compare • Conclusion • Contrast • Discover • Dissect • Distinguish • Divide • Examine • Function • Inference • Inspect • List • Motive • Relationships • Simplify • Survey • Take part in • Test for • Theme	• Agree • Appraise • Assess • Award • Choose • Compare • Conclude • Criteria • Criticize • Decide • Deduct • Defend • Determine • Disprove • Estimate • Evaluate • Explain • Importance • Influence • Interpret • Judge • Justify • Mark • Measure • Opinion • Perceive • Prioritize • Prove • Rate • Recommend • Rule on • Select • Support • Value	• Adapt • Build • Change • Choose • Combine • Compile • Compose • Construct • Create • Delete • Design • Develop • Discuss • Elaborate • Estimate • Formulate • Happen • Imagine • Improve • Invent • Make up • Maximize • Minimize • Modify • Original • Originate • Plan • Predict • Propose • Solution • Solve • Suppose • Test • Theory

Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing, Abridged Edition. Boston, MA: Allyn and Bacon.

Action Words for Bloom's Taxonomy					
Knowledge	**Understand**	**Apply**	**Analyze**	**Evaluate**	**Create**
define	explain	solve	analyze	reframe	design
identify	describe	apply	compare	criticize	compose
describe	interpret	illustrate	classify	evaluate	create
label	paraphrase	modify	contrast	order	plan
list	summarize	use	distinguish	appraise	combine
name	classify	calculate	infer	judge	formulate
state	compare	change	separate	support	invent
match	differentiate	choose	explain	compare	hypothesize
recognize	discuss	demonstrate	select	decide	substitute
select	distinguish	discover	categorize	discriminate	write
examine	extend	experiment	connect	recommend	compile
locate	predict	relate	differentiate	summarize	construct
memorize	associate	show	discriminate	assess	develop
quote	contrast	sketch	divide	choose	generalize
recall	convert	complete	order	convince	integrate
reproduce	demonstrate	construct	point out	defend	modify
tabulate	estimate	dramatize	prioritize	estimate	organize
tell	express	interpret	subdivide	find errors	prepare
copy	identify	manipulate	survey	grade	produce
discover	indicate	paint	advertise	measure	rearrange
duplicate	infer	prepare	appraise	predict	rewrite
enumerate	relate	produce	break down	rank	role-play
listen	restate	report	calculate	score	adapt
observe	select	teach	conclude	select	anticipate
omit	translate	act	correlate	test	arrange
read	ask	administer	criticize	argue	assemble
recite	cite	articulate	deduce	conclude	choose
record	discover	chart	devise	consider	collaborate
repeat	generalize	collect	diagram	critique	collect
retell	give examples	compute	dissect	debate	devise
visualize	group	determine	estimate	distinguish	express
	illustrate	develop	evaluate	editorialize	facilitate
	judge	employ	experiment	justify	imagine
	observe	establish	focus	persuade	infer
	order	examine	illustrate	rate	intervene
	report	explain	organize	weigh	justify
	represent	interview	outline		make
	research	judge	plan		manage
	review	list	question		negotiate
	rewrite	operate	test		originate
	show	practice			propose
	trace	predict			reorganize
	transform	record			report
		schedule			revise
		simulate			schematize
		transfer			simulate
		write			solve
					speculate
					structure
					support
					test
					validate

4. Syllabus copy

Unit– I	**Introduction to R:** Handling Packages in R, Getting Started with R, Working with Directory, Data Types in R, Commands for Data Exploration Loading and Handling Data in R:Challenges of Analytical Data Processing, Expression, Variables and Functions, Missing Values Treatment in R, Using the 'as' Operator to Change the Structure of Data, Vectors, Matrices, Factors, List, Aggregating and Group Processing of a Variable, Simple Analysis Using R, Methods for Reading Data, Comparison of R GUIs for Data Input.
Unit– II	**Descriptive Statistics:** Using Statistics, Percentiles and Quartiles, Measures of Central Tendency, Measures of Variability, Grouped Data and the Histogram, Skewness and Kurtosis, Relations between the Mean and the Standard Deviation, Methods of Displaying Data, Exploratory Data Analysis.
Unit– III	**Linear Regression using R:** Introduction, Model Fitting, Linear Regression Assumptions of Linear Regression, Validating Linear Assumption
Unit– IV	**Logistic Regression using R:** Introduction, Introduction to Generalized Linear Models, Logistic Regression, Binary Logistic Regression, Diagnosing Logistic Regression, Multinomial Logistic Regression Models.
Unit– V	**Decision Tree:** Introduction, Decision Tree Representation in R, Appropriate Problems for Decision Tree Learning, Basic Decision Tree Learning Algorithm, Measuring Features, Hypothesis Space Search in Decision Tree Learning, Inductive Bias in Decision Tree Learning, Why Prefer Short Hypotheses, Issues in Decision Tree Learning.

4.1 References (Text books/websites/Journals)

TEXT BOOK:

1. Seema Acharya - "Data Analytics Using R" ,Jan 01, 2018, Seema Acharya-MC GRAW HILL INDIA(2018).
2. Aczel−Sounder pandian: "Complete Business Statistics" 7th Edition Complete Business Statistics, Seventh Edition McGraw−Hill Primis.

REFERENCESBOOKS:

1. ROBERT I. KABACOFF "R in ActionData analysis and graphics with R" Manning Publications Co2011

Journals with min 5 ref paper for literature study

1. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevska O; written on behalf of AME Big-Data Clinical Trial Collaborative Group. Predictive analytics with gradient boosting in clinical medicine. Ann Transl

2. Besharati, E., Naderan, M. & Namjoo, E. LR-HIDS: logistic regression host-based intrusion detection system for cloud environments. J Ambient Intell Human Comput 10, 3669–3692 (2019). https://doi.org/10.1007/s12652-018-1093-8

3. Kliestik, T., Kliestikova, J., Kovacova, M., Svabova, L., Valaskova, K., Vochozka, M., & Olah, J. (2018). Prediction of financial health of business entities in transition economies. New York: Addleton Academic Publishers.

4. Xilei Zhao, Xiang Yan, Alan Yu, Pascal Van Hentenryck, Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models, Travel Behaviour and Society, Volume 20, 2020,Pages 22-35, ISSN 2214-367X, https://doi.org/10.1016/j.tbs.2020.02.003

5. Piryonesi, S.M. and El-Diraby, T.E., 2020. Role of data analytics in infrastructure asset management: Overcoming data size and quality problems. *Journal of Transportation Engineering, Part B: Pavements*, *146*(2), p.04020022.

5. Session plan

S. No	Topics	Sub-Topic	No. Of Lectures Required	Books	Methods
UNIT - I					
1	Introduction to R	Handling Packages in R, Getting Started with R,	L1	T1, R1	M1, M2, M4
2		Working with Directory, Data Types in R, Commands for Data	L2	T1, R1	M1, M2, M4
3		Exploration Loading and Handling Data in R:	L3	T1, R1	M1, M2, M4
4		Challenges of Analytical Data Processing, Expression, Variables and	L4	T1, R1	M1, M2, M4
5		Functions Using the 'as' Operator to Change the Structure of Data	L5	T1, R1	M1, M2, M4
		Vectors, Matrices, Lists	L6, L7	T1, R1	M1, M2, M4
5	Simple Analysis Using R	Aggregating and Group Processing of a Variable,	L8	T1, R1	M1, M2, M4
6		Simple Analysis Using R,	L9	T1, R1	M1, M2, M4
7		Methods for Reading Data,	L10	T1, R1	M1, M2, M4
8		Comparison of R GUIs for Data Input	L11, L12	T1, R1	M1, M2, M4
UNIT - II					
8	Descriptive Statistics	Using Statistics, Percentiles and Quartiles,	L13	T1, R1	M1, M2, M4
9		Measures of Central Tendency,	L14	T1, R1	M1, M2, M4
10		Measures of Variability,	L15	T1, R1	M1, M2, M4
11		Grouped Data and the Histogram, Skewness and Kurtosis,	L16	T1, R1	M1, M2, M4

12	Exploratory Data Analysis.	Relations between the Mean and the Standard Deviation,	L17	T1, R1	M1, M2, M4
13		Methods of Displaying Data	L18	T1, R1	M1, M2, M4
14		Exploratory Data Analysis.	L19, L20	T1, R1	M1, M2, M4
UNIT -III					
21	Linear Regression using R	Introduction	L21	T1, R1	M1, M2, M4
22		Model Fitting,	L22	T1, R1	M1, M2, M4
23		Linear Regression	L23, L24	T1, R1	M1, M2, M4
26	Validating Linear Assumption	Assumptions of Linear Regression	L25	T1, R1	M1, M2, M4
27		Validating Linear Assumption	L26, 27, 28	T1, R1	M1, M2, M4
UNIT -IV					
31	Logistic Regression using R	Introduction	L29	T1, R1	M1, M2, M4
32		Generalized Linear Models	L30	T1, R1	M1, M2, M4
33		Logistic Regression	L31, L32	T1, R1	M1, M2, M4
38	Diagnosing Logistic Regression	Introduction to Binary Logistic Regression	L33	T1, R1	M1, M2, M4
39		Diagnosing Logistic Regression	L34, L35	T1, R1	M1, M2, M4
40		Multinomial Logistic Regression Models.	L36, L37	T1, R1	M1, M2, M4
UNIT -V					
42	Decision Tree	Introduction	L38	T1, R1	M1, M2, M4
43		Decision Tree Representation in R	L39	T1, R1	M1, M2, M4
44		Appropriate Problems for Decision Tree Learning	L40	T1, R1	M1, M2, M4
46	Decision tree algorithm	Basic Decision Tree Learning Algorithm	L41, L42, L43	T1, R1	M1, M2, M4
47		Measuring Features, Hypothesis Space Search in Decision Tree Learning	L44, L45	T1, R1	M1, M2, M4
48		Why Prefer Short Hypotheses, Issues in Decision Tree Learning.	L46, L47, L48	T1, R1	M1, M2, M4

Total Classes =48

INDIVIDUAL TIME TABLE (Dr. C.N.Ravi)

Day/Hour	1 hour	2 hour	3 hour	4 hour	5 hour	6 hour	7 hour
Mon			CSE A		CSE C		CSE C
Tue							
Wed		CSE C		CSE C	CSE A		
Thu	CSE A			CSE C			
Fri		CSE A				CSE A	

6. Session Execution Log

S No	Unit	Scheduled complete date	Completed date	Remarks
1	**I**	21/08/2023	14/09/2023	COMPLETED
2	**II**	15/09/2023	30/09/2023	COMPLETED
3	**III**	01/10/2023	22/10/2023	COMPLETED
4	**IV**	24/10/2023	15/11/2023	COMPLETED
5	**V**	18/11/2023	23/12/2023	COMPLETED

7. Lecture Notes

Attached

8. Assignment Questions along with sample Assignment Script

Assignment-1

1. **Read** in the dataset ("Cereals.csv"), look at the first few rows with head and inspect the data types of the variables in the dataframe with str. Write the R Code for the following.
   (i). Add a new variable to the dataset called 'totalcarb', which is the sum of 'carbo' and 'sugars'.
   (ii). How many cereals in the dataframe are 'hot' cereals? Hint: take an appropriate subset of the data, and then count the number of observations in the subset.
   (iii). How many unique manufacturers are included in the dataset?
   (iv). Take a subset of the dataframe with only the Manufacturer 'K' (Kellogg's)
   (v). Take a subset of the dataframe of all cereals that have less than 80 calories, AND have more than 20 units of vitamins. (CO1)

2. **What** is EDA, Write down the Exploratory Data Analysis(EDA) steps involved in Data Science. (CO2)
3. **Explain** the IQR and Outlier of IQR with Example. (CO2)
4. **What** is Model Fitting.(CO3)
5. **Write** down the DataStructure Data types used in R with examples. (CO1)

Assignment-2

DAR Assignment -2

1. Explain the Process of fitting a Linear Regression Model R.
2. Evaluate the effectiveness of using multinomial logistic regression models for predicting multiple outcomes.
3. Describe the Role Maximum likelihood estimation in Logistic regression and how it is used to fir the model.
4. Discuss the Issues in Decision tree construction.
5. Draw the Decision Tree and find out the “entropy” and “information gain” for this dataset.

Outlook	Temperature	Humidity	Windy	PlayTennis
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

9. Mid exam Question Papers along with sample Answer Scripts

III. B. TECH- I-SEM-I MID EXAMINATION ***Date:*** **Time:** ***09:09:2023/10:00-11:30 AM***
Subject: Data Analytics Using R (CS512PE) **Branch: CSE**
Marks: 25 M

Note: Question paper contains two parts, Part - A and Part - B.
Part-A is compulsory which carries 10 marks. Answer all questions in part-A.
Part-B answer 3 questions, each question carries 5 marks.

PART-A **5x2=10**

1 Q. List features of R programming Language?	(CO1)
2 Q. Name few packages used for data management in R	(CO1)
3 Q. What are the objectives of an EDA process	(CO1)
4 Q. What is the use of lm() function in R	(CO2)
5 Q. What is R2 statistic	(CO2)

PART-B **3X5=15**

6 Q. Write about Data types and objects available in R	(CO1)
(OR)	
7. Q. Explain how to handle missing values using R functions	(CO2)
8. Q. Elaborate on how to use MySQL database in R	(CO1)
(OR)	
9. Q. Write the measures used for computing centrality and dispersion of data	(CO2)
9. Q. Illustrate with an example different methods of data visualizations	(CO4)
(OR)	
10. Q. Discuss the relation between mean and standard deviation	(CO2)

CMR ENGINEERING COLLEGE
UGC AUTONOMOUS
(Approved by AICTE - New Delhi. Affiliated to JNTUH and Accredited by NAAC & NBA)
Kandlakoya (V), Medchal (M), Medchal - Malkajgiri (D)-501401

III. B. TECH- I-SEM-II MID EXAMINATION ***Date:*** **Time: Date: 30/12/2023**
Subject: Data Analytics Using R (CS512PE) **Branch: CSE A & C**
Marks: 25 M

Note: Question paper contains two parts, Part - A and Part - B.
Part-A is compulsory which carries 10 marks. Answer all questions in part-A.
Part-B answer 3 questions, each question carries 5 mark

Part –A

1. What is the importance of the F-test in a linear model? (CO1)
2. What is the difference between residual and goodness-of-fit tests ?.(CO1)
3. What is the difference between nlm() and optim() functions ?(CO2)
4. Which packages build decision trees in R ?(CO2)
5. What do you mean by the disjunction form ?(CO3)

Part –B

Answer any 3 Questions: Marks: 3X5=15M

6. What is model Fitting? Explain various models and their commands in R. (CO1)
7. Create a table with a 'pizza' column that stores the information that is necessary to implement multinomial logistics regression. After placing the information, implement multinomial logistics regression on this table?(CO1)
8. Which function implements the GLM model in R? Explain with an example and syntax (CO2)
9. Create a dataset that contains the features of apples. Now find out the "entropy" and "information gain" for this dataset. Also, find out the best feature of the apple dataset.(CO3)
10. Create a dataset and generate the decision tree for it using the ctree() function?(CO2)
11. Take any inbuilt dataset from R and explain pruning in this dataset ?(CO2)

Scheme of Evaluation

MID-I

Scheme of Evaluation

Question. No.	Theory	Marks	Total
Part -A			
1	4 features	0.5*4	2
2	4 package names in R	0.5*4	2
3	2 objectives	2	2
4	lm() function use	2	2
5	Definition and Formula of R^2	1*2	2
Part -B			
6	Primary datatypes	2	5
	Objects and derived datatypes syntax and examples	3	
7	Missing values representation in R	2	5
	Function in R and removal techniques	3	
8	Importing Mysql and initializing database in R	2.5	5
	Some DML commands examples	2.5	
9	Measures of centrality	2.5	5
	Measures of dispersion	2.5	
10	Data visualization techniques	2.5	5
	ggplot functions in R	2.5	
11	Explain mean and std dev	2.5	5
	Provide their relationship	2.5	

MID-II

Scheme of Evaluation

Question. No.	Theory	Marks	Total
Part -A			
1	Explanation	2	2
2	2 difference	1*2	2
3	2 difference	1*2	2
4	Explanation	2	2
5	Explanation	2	2
Part -B			
6	Model Fitting	2	5
	Various Models	3	
7	Creating table	2	5
	MLE Implementation	3	
8	Function Implementation	2.5	5
	Syntax and Example	2.5	
9	Creating table	2	5
	Entropy & IG	3	
10	Decision tree construction process	2.5	5
	Ctree	2.5	
11	Pruning process	5	5

11. Mappings of Cos with Pos and PSOs

Course Outcomes	Relationship of Course outcomes to Program Outcomes (PO AVG)													
CO/PO	**PO1**	**PO2**	**PO3**	**PO4**	**PO5**	**P06**	**PO7**	**P08**	**P09**	**PO10**	**PO11**	**PO12**	**PSO1**	**PSO2**
CO1	3	1	2	1	3					2	2	1	2	2
CO2	2	2	2	3	3	2				2	1	1	2	2
CO3	2	1	3	3	2	1							1	1
CO4	1	1	3	3	2								1	1
CO5	2	1	3	3	1					2	1	1	2	2
Average	**2**	**1**	**3**	**3**	**2**	**1**	**0**	**0**	**0**	**1**	**1**	**1**	**2**	**2**

10. COs, POs, PSOs Justification

Justification:

CO1: Understanding the building blocks of R-Programming **(Understanding)**
PO1 is strongly correlated: Learning R programming provides a student with an ability to apply the knowledge of computing to real world problem
PO2 is weekly correlated: Provides some basics of finding solution for computational problem
PO3 is moderately correlated: It provides ability to design, implement, and evaluate a computer-based system
PO4 is weekly correlated: Some ability to design and conduct experiments, as well as to analyze and interpret data
PO5 is strongly correlated: R programming is modern programming language for data analytics
PO10 is moderately correlated: An ability to communicate effectively with a range of audiences through analyzed data and visualizations
PO11 is moderately correlated: To some extents helps in project management
PO12 is weekly correlated: R Programming is lifelong learning skill by engaging oneself in developing projects in R
PSO1 is moderately correlated: Learning R programming skills gives an ability of developing projects
PSO2 is moderately correlated: Learning some real case studies of data analytics in R help in research skills

CO2: Apply critical R-Programming concepts to handle the data **(Applying)**
PO1 is moderately correlated: Learning R programming provides a student with an ability to apply the knowledge of computing to real world problem
PO2 is moderately correlated: Provides some basics of finding solution for computational problem
PO3 is moderately correlated: It provides ability to design, implement, and evaluate a computer-based system
PO4 is strongly correlated: Provides ability to design and conduct experiments, as well as to analyze and interpret data

PO5 is strongly correlated: R programming is a modern programming language for data analytics
PO6 is moderately correlated: Gives an ability to analyze the local and global impact of computing on individuals, organizations, and society
PO10 is moderately correlated: An ability to communicate effectively with a range of audiences through analyzed data and visualizations
PO11 is weekly correlated: To some extents helps in project management
PO12 is weekly correlated: R Programming is lifelong learning skill by engaging oneself in developing projects in R
PSO1 is moderately correlated: It provides ability to design, implement, and evaluate a computer-based system
PSO2 is moderately correlated: Learning some real case studies of data analytics in R help in research skills

CO3: Apply statistical concepts on real data (Applying)
PO1 is moderately correlated: Learning basic statistics provides a student with an ability to apply the knowledge of computing to real world problem
PO2 is weekly correlated: Provides some basics of finding statistical analysis for computational problems
PO3 is strongly correlated: It provides ability to design, implement, and evaluate a computer-based system
PO4 is strongly correlated: Provides ability to design and conduct experiments, as well as to analyze and interpret data using the knowledge of statistics
PO5 is moderately correlated: Learn modern Statistical tools like R
PO6 is weekly correlated: Gives some ability to analyze the local and global impact of computing on individuals, organizations, and society
PSO1 is weekly correlated: Statistical analysis provides some ability to design, implement, and evaluate a computer-based system
PSO2 is weekly correlated: Studying real case studies of statistical data analytics help in improving research skills

CO4: Analyze logistic and linear regression models on real data (Analyze)
PO1 is weekly correlated: Learning predictive and classification models provides a student with an ability to apply the knowledge of computing to real world problem
PO2 is weekly correlated: Provides some basics of learning to solve computational problems
PO3 is strongly correlated: It provides the student with an ability to design, implement, and evaluate a learning system
PO4 is strongly correlated: Provides ability to design and conduct experiments, as well as to analyze and interpret data using the knowledge of predictive models
PO5 is moderately correlated: These are modern learning techniques useful for real world applications
PSO1 is weekly correlated: Gives some ability to design, implement, and evaluate a AI system
PSO2 is weekly correlated: Application of these techniques help in improving research skills

CO5: Create decision trees to classify the data (Create)
PO1 is moderately correlated: Decision tree models provides a student with an ability to apply the knowledge of computing to real world problem
PO2 is weekly correlated: Provides some problem-solving skills
PO3 is strongly correlated: It provides the student with an ability to design, implement, and evaluate a decision support system
PO4 is strongly correlated: Provides ability to design and conduct experiments, as well as to analyze and interpret data using decision trees
PO5 is weekly correlated: Modern technique
PO10 is moderately correlated: An ability to communicate effectively with a range of audiences through tree visualizations
PO11 is weekly correlated: To some extents helps in project management
PO12 is weekly correlated: Lifelong learning skills by engaging oneself in developing projects
PSO1 is moderately correlated: Gives some ability to design, implement, and evaluate a AI system
PSO2 is moderately correlated: Application of these techniques help in improving research skills

13. Attainment of COs, POs and PSOs

14. University question papers or question bank.

Previous Year QP papers

CMR ENGINEERING COLLEGE: : HYDERABAD
UGC AUTONOMOUS
III–B.TECH–I–Semester End Examinations (Regular) - December- 2022
DATA ANALYTICS USING R
(CSE)

[Time: 3 Hours] **[Max. Marks: 70]**

Note: This question paper contains two parts A and B.
Part A is compulsory which carries 20 marks. Answer all questions in Part A.
Part B consists of 5 Units. Answer any one full question from each unit. Each question carries 10 marks.

PART-A **(20 Marks)**

1.	a)	How can the default path to package library be changed in R?	[2M]
	b)	List out two IDEs for R.	[2M]
	c)	What are mean and median with a neat example?	[2M]
	d)	What are the advantages of using data visualization?	[2M]
	e)	Give the general equation for computing linear regression?	[2M]
	f)	What is the syntax of lm() function?	[2M]
	g)	What is a three-way contingency table?	[2M]
	h)	What are the major diagnostic functions of the 'LogisticDx' package?	[2M]
	i)	Name the packages used to build decision trees in R?	[2M]
	j)	List out the names of learning algorithms that create a decision tree.	[2M]

PART-B **(50 Marks)**

2.	a)	Explain RSQLite package.	[5M]
	b)	Explain the commands using R: summary (), str (), head (), tail (), view (), edit ()	[5M]
		OR	
3.		Create a dataset, 'Watch' and store the information about watches of four different companies. Explain all the steps of simple analytical data processing from input to output on this dataset.	[10M]
4.	a)	What are the data frames? Write its significance in R-Language?	[5M]
	b)	Explain the graphical techniques used by Exploratory Data Analysis using R.	[5M]
		OR	
5.		What is bar chart? Discuss the various types of bar charts using R?	[10M]
6.		Compare and Contrast Multiple R-squared and Adjusted R-squared.	[10M]
		OR	
7.		What is model Fitting? Explain various models and their commands in R.	[10M]

8. Create a table with a 'pizza' column that stores the information that is necessary to implement multinomial logistics regression. After placing the information, implement multinomial logistics regression on this table. [10M]

OR

9. a) Explain binary logistic regression with a single categorical variable. [5M]
   b) Explain about likelihood function. [5M]

10. Create a dataset that contains the features of apples. Now find out the "entropy" and "information gain" for this dataset. Also, find out the best feature of the apple dataset. [10M]

OR

11. Write and explain ID3 decision tree construction algorithm. [10M]

****

Code No.: CS512PE | R20 | H.T.No. | 8 | R

CMR ENGINEERING COLLEGE: : HYDERABAD
UGC AUTONOMOUS
III–B.TECH–I–Semester End Examinations (Supply) - May- 2023
DATA ANALYTICS USING R
(CSE)

[Time: 3 Hours] **[Max. Marks: 70]**

Note: This question paper contains two parts A and B.
Part A is compulsory which carries 20 marks. Answer all questions in Part A.
Part B consists of 5 Units. Answer any one full question from each unit. Each question carries 10 marks.

PART-A **(20 Marks)**

1. a)	Identify the various data types in R.	[2M]
b)	Compare and contrast the various R GUIs for data input.	[2M]
c)	Difference between variance and standard deviation.	[2M]
d)	Define measures of central tendency.	[2M]
e)	Define linear regression in statistics.	[2M]
f)	Identify the assumptions of linear regression and explain why they are important.	[2M]
g)	Define binary logistic regression and its application.	[2M]
h)	Analyze the goodness of fit for a binary logistic regression model.	[2M]
i)	How to Identify the appropriate problems for decision tree learning.	[2M]
j)	Analyze the representation of decision tree in R.	[2M]

PART-B **(50 Marks)**

2.a)	Evaluate the effectiveness of the 'as' operator in changing the structure of data in R.	[7M]
b)	How do you install and load packages in R?	[3M]
	OR	
3.	Using R, create a matrix and perform basic arithmetic operations on it.	[10M]
4.	Compare and contrast the use of Mean, Median, and Mode as measures of central tendency in different scenarios.	[10M]
	OR	
5.a)	Evaluate the effectiveness of using Histograms versus Box plots for displaying data in statistics.	[5M]
b)	What is the difference between skewed and symmetric data?	[5M]
6.	Evaluate the effectiveness of different methods for model validation in linear regression.	[10M]
	OR	
7.a)	Explain the difference between correlation and linear regression.	[5M]
b)	Explain the process of fitting a linear regression model in R.	[5M]
8.	Describe the role of maximum likelihood estimation in Logistic Regression and how it is used to fit the model.	[10M]
	OR	
9.	Evaluate the effectiveness of using multinomial Logistic Regression models for predicting multiple outcomes.	[10M]
10.a)	Analyze the impact of data preprocessing on Decision Tree learning.	[7M]
b)	What is inductive biasing in decision tree learning?	[3M]
	OR	
11.	Analyze the advantages and disadvantages of using different measures of impurity, such as Entropy and Gini index, in Decision Tree learning.	[10M]

****

Question banks:

UNIT-1

1. What is R?
2. What is the predecessor of R?
3. What is the fundamental data type of R?
4. What is the disadvantage of using R in enterprise-level large-scale solutions?
5. How to locate an RScript file in a typical file system?
6. What is R markdown and how is it different from word documentation?
7. Name a few packages used for data management in R.
8. Name a few packages used for data visualisation in R.
9. Name a few packages used for developing data produces in R.
10. Name a few packages used for data modelling and simulation in R.

Long Answer:

1. How can the default path to package library be changed in R?
2. What is the command to check and install the “dplyr” package?
3. How can we install multiple packages in R?
4. What are the differences between the head() and tail() commands in R?
5. What does the data() function help with?
6. What is nrow() function?
7. What is analytical data processing?
8. List the challenges of analytical data processing.
9. What are the common steps of analytical data processing?
10. What is the na.exclude() function?

UNIT-2

Short Questions

1. List the differences between the head() and tail() functions?
2. What is EDA?
3. Differentiate between invalid values and outliers.
4. How are missing values treated in R?
5. What is data visualisation?
6. How to calculate a data range?
7. How to find a mode value?
8. Give contrast of mean and median.
9. What is density plot?
10. What is histogram?

Long Questions

1. Explain the reason to use the trim parameter.
2. Create a histogram by filling the bar with 'blue' colour.
3. What is a bar chart? Describe the types of bar charts.
4. Create a horizontal bar chart.
5. Differentiate between a group and stacked bar chart.
6. Create and place a legend in bar chart

UNIT –III

1. What is a Linear Regression?
2. Can you list out the critical assumptions of linear regression?
3. What is Heteroscedasticity?
4. What is the primary difference between R square and adjusted R square?
5. Can you list out the formulas to find RMSE and MSE?
6. Can you name a possible method of improving the accuracy of a linear regression model?
7. What are outliers? How do you detect and treat them?
8. How do you interpret a Q-Q plot in a linear regression model?
9. What is the importance of the F-test in a linear model?
10. What are the disadvantages of the linear regression model?

Long Questions

1. What is model fitting?
2. What is the general equation for computing linear regression?
3. What is a response and predictor variable?
4. What is the syntax of lm() function?
5. What is a residual?
6. What is leverage?
7. What is Cook's distance?
8. What is homoscedasticity?
9. How to find standard error?
10. How to plot a scatterplot?

UNIT - 4

Short Questions

1. What is GLM regression? What are its components?
2. What are the applications of logistic regression?
3. What are independent and dependent variables in regression?
4. What is the difference between the logistic and logit functions?
5. What is the difference between nlm() and optim() functions?
6. What is the difference between Pearson and deviance residuals?
7. What is the difference between residual and goodness-of-fit tests?

Long Questions

1. Which function implements the GLM model in R? Explain with an example and syntax.
2. Explain the nlm() function with syntax and an example.
3. Explain the optim() function with syntax and an example.
4. Explain the mle() function with syntax and an example.
5. Explain binary logistic regression with a single categorical variable.
6. Explain binary logistic regression with a contingency table.
7. Explain binary logistic regression with a covariate variable.
8. Explain the multinom() function with syntax and an example.
9. Create a table with an 'employee' column that stores the necessary information including each employee's performance scores. Implement logistic regression to check whether an employee is eligible for promotion or not based on his/her performance score. Also, implement the mle() function for defining the maximum likelihood estimation.
10. Create a table with a 'person' column that stores the information like name, age, gender, annual income and other. Implement the binary logistic regression with single categorical and three-way contingency table after placing the required information on the table.
11. Create a table with a 'pizza' column that stores the information that is necessary to implement multinomial logistics regression. After placing the information, implement multinomial logistics regression on this table.

UNIT –V

.

Short Questions

1. What is the role of decision trees in machine learning? How many types of trees are used in machine learning?
2. Write about the packages, ‘rpart’ and ‘party’.
3. What is the difference between CTree and ctree() in R?
4. What is the decision-tree learning algorithm?
5. What are the applications of the decision-tree learning algorithm?
6. What is hypothesis space search? List its steps.
7. What are the methods to resolve “the missing attributes value problem” in the decision tree?

Long Questions

1. Think of a problem statement and represent it using a decision tree.
2. Explain the packages data.tree, entropy and information gain with examples.
3. Explain “Occam’s razor”.
4. What is pruning? Why it is used in a decision tree?
5. Explain the prune() function with syntax and an example.
6. Create a dataset and generate the decision tree for it using the ctree() function.
7. Create a dataset that contains attribute-value pairs. Generate the decision tree for it using the ctree() function.
8. Create a dataset that contains attribute-value pairs. Generate the decision tree for it using the ctree() function.
9. Create a dataset that contains discrete values. Generate the decision tree for it using the ctree() function.
10. Create a dataset that contains the data in disjunction form. Generate the decision tree for it using the ctree() function.
11. Take any inbuilt dataset from R and explain pruning in this dataset.
12. Create a dataset that contains the features of apples. Now find out the “entropy” and “information gain” for this dataset. Also, find out the best feature of the apple dataset.

Previous Year QP papers

CMR ENGINEERING COLLEGE: : HYDERABAD
UGC AUTONOMOUS
III–B.TECH–I–Semester End Examinations (Regular) - December- 2022
DATA ANALYTICS USING R
(CSE)

[Time: 3 Hours] **[Max. Marks: 70]**

Note: This question paper contains two parts A and B.
Part A is compulsory which carries 20 marks. Answer all questions in Part A.
Part B consists of 5 Units. Answer any one full question from each unit. Each question carries 10 marks.

PART-A **(20 Marks)**

1.	a)	How can the default path to package library be changed in R?	[2M]
	b)	List out two IDEs for R.	[2M]
	c)	What are mean and median with a neat example?	[2M]
	d)	What are the advantages of using data visualization?	[2M]
	e)	Give the general equation for computing linear regression?	[2M]
	f)	What is the syntax of lm() function?	[2M]
	g)	What is a three-way contingency table?	[2M]
	h)	What are the major diagnostic functions of the 'LogisticDx' package?	[2M]
	i)	Name the packages used to build decision trees in R?	[2M]
	j)	List out the names of learning algorithms that create a decision tree.	[2M]

PART-B **(50 Marks)**

2. a) Explain RSQLite package. [5M]
   b) Explain the commands using R: summary (), str (), head (), tail (), view (), edit () [5M]

OR

3. Create a dataset, 'Watch' and store the information about watches of four different companies. Explain all the steps of simple analytical data processing from input to output on this dataset. [10M]

4. a) What are the data frames? Write its significance in R-Language? [5M]
   b) Explain the graphical techniques used by Exploratory Data Analysis using R. [5M]

OR

5. What is bar chart? Discuss the various types of bar charts using R? [10M]

6. Compare and Contrast Multiple R-squared and Adjusted R-squared. [10M]

OR

7. What is model Fitting? Explain various models and their commands in R. [10M]

8. Create a table with a 'pizza' column that stores the information that is necessary to implement multinomial logistics regression. After placing the information, implement multinomial logistics regression on this table. [10M]

OR

9. a) Explain binary logistic regression with a single categorical variable. [5M]
   b) Explain about likelihood function. [5M]

10. Create a dataset that contains the features of apples. Now find out the “entropy” and “information gain” for this dataset. Also, find out the best feature of the apple dataset. [10M]

OR

11. Write and explain ID3 decision tree construction algorithm. [10M]

****

Code No.: CS512PE | R20 | H.T.No. | 8 | R

CMR ENGINEERING COLLEGE: : HYDERABAD
UGC AUTONOMOUS
III–B.TECH–I–Semester End Examinations (Supply) - May- 2023
DATA ANALYTICS USING R
(CSE)

[Time: 3 Hours] **[Max. Marks: 70]**

Note: This question paper contains two parts A and B.
Part A is compulsory which carries 20 marks. Answer all questions in Part A.
Part B consists of 5 Units. Answer any one full question from each unit. Each question carries 10 marks.

PART-A **(20 Marks)**

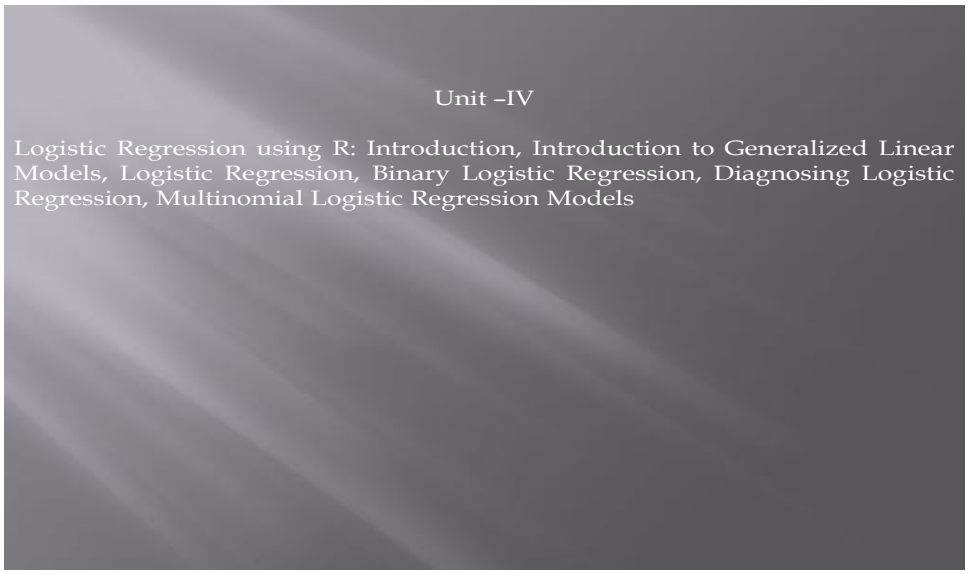
1.	a)	Identify the various data types in R.	[2M]
	b)	Compare and contrast the various R GUIs for data input.	[2M]
	c)	Difference between variance and standard deviation.	[2M]
	d)	Define measures of central tendency.	[2M]
	e)	Define linear regression in statistics.	[2M]
	f)	Identify the assumptions of linear regression and explain why they are important.	[2M]
	g)	Define binary logistic regression and its application.	[2M]
	h)	Analyze the goodness of fit for a binary logistic regression model.	[2M]
	i)	How to Identify the appropriate problems for decision tree learning.	[2M]
	j)	Analyze the representation of decision tree in R.	[2M]

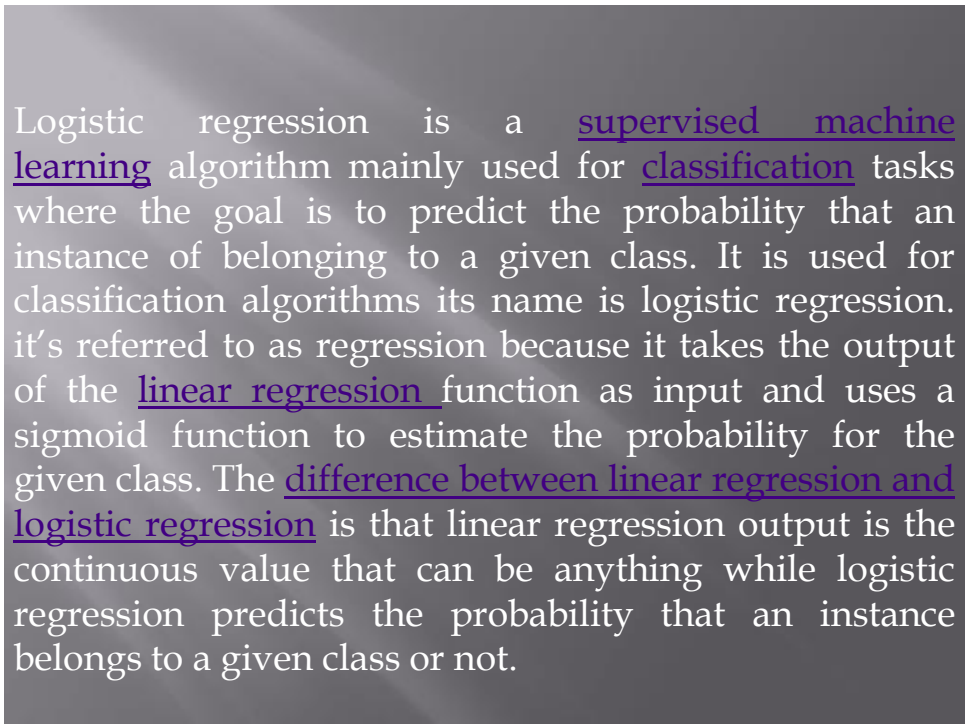
PART-B **(50 Marks)**

2.	a)	Evaluate the effectiveness of the 'as' operator in changing the structure of data in R.	[7M]
	b)	How do you install and load packages in R?	[3M]
		OR	
3.		Using R, create a matrix and perform basic arithmetic operations on it.	[10M]
4.		Compare and contrast the use of Mean, Median, and Mode as measures of central tendency in different scenarios.	[10M]
		OR	
5.	a)	Evaluate the effectiveness of using Histograms versus Box plots for displaying data in statistics.	[5M]
	b)	What is the difference between skewed and symmetric data?	[5M]
6.		Evaluate the effectiveness of different methods for model validation in linear regression.	[10M]
		OR	
7.	a)	Explain the difference between correlation and linear regression.	[5M]
	b)	Explain the process of fitting a linear regression model in R.	[5M]
8.		Describe the role of maximum likelihood estimation in Logistic Regression and how it is used to fit the model.	[10M]
		OR	
9.		Evaluate the effectiveness of using multinomial Logistic Regression models for predicting multiple outcomes.	[10M]
10.	a)	Analyze the impact of data preprocessing on Decision Tree learning.	[7M]
	b)	What is inductive biasing in decision tree learning?	[3M]
		OR	
11.		Analyze the advantages and disadvantages of using different measures of impurity, such as Entropy and Gini index, in Decision Tree learning.	[10M]

****

15.Power Point Presentations (PPTs)





Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value.

It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).

The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

Logistic Function (Sigmoid Function):

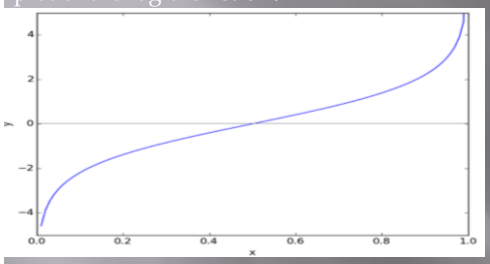
The sigmoid function is a mathematical function used to map the predicted values to probabilities.

It maps any real value into another value within a range of 0 and 1. o The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form.

The S-form curve is called the Sigmoid function or the logistic function.

In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

The **logit function** is simply the logarithm of the odds: $logit(x) = log(x / (1 - x))$. Here is a plot of the logit function:

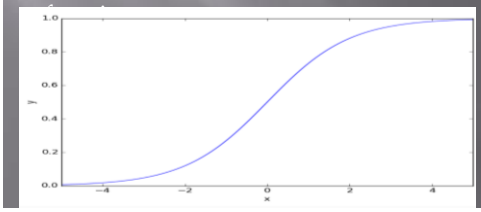


The value of the logit function heads towards infinity as p approaches 1 and towards negative infinity as it approaches 0.

The logit function is useful in analytics because it maps probabilities (which are values in the range [0, 1]) to the full range of real numbers

The inverse of the logit function is the **sigmoid** function. That is, if you have a probability p, $sigmoid(logit(p)) = p$. The sigmoid function maps arbitrary real values back to the range [0, 1]. The larger the value, the closer to 1 you'll get.

The formula for the sigmoid function is $\sigma(x) = 1/(1 + exp(-x))$. Here is a plot of the



The sigmoid might be useful if you want to transform a real valued variable into something that represents a probability

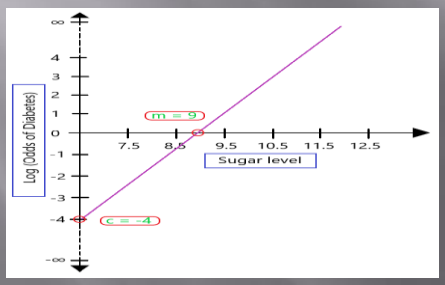
Type of Logistic Regression:
On the basis of the categories, Logistic Regression can be classified into three types:
Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High"

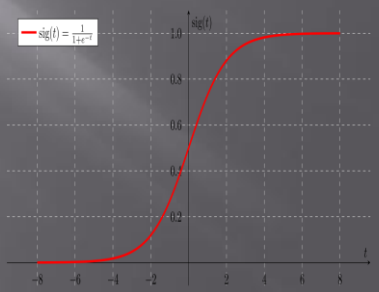
Binary response & regression models

We have a **binary** (dichotomous) response variable Y defined as

Y = 1 if "success" ("yes")
0 if "failure" ("no")

We want to model the **probability** Π that Y=1





Logit function to Sigmoid Function

Logistic Regression can be expressed as,

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta 0 + \beta 1X$$

where p(x)/(1-p(x)) is termed odds, and the left-hand side is called the logit or log-odds function. The odds are the ratio of the chances of success to the chances of failure. As a result, in Logistic Regression, a linear combination of inputs is translated to log(odds), with an output of 1.

The following is the inverse of the aforementioned function

$$p(X) = \frac{e^{\beta 0+\beta 1X}}{1 + e^{\beta 0+\beta 1X}}$$

This is the Sigmoid function, which produces an S-shaped curve. It always returns a probability value between 0 and 1. The Sigmoid function is used to convert expected values to probabilities. The function converts any real number into a number between 0 and 1. We utilize sigmoid to translate predictions to probabilities in machine learning.

The mathematically sigmoid function can be,

$$f(x) = \frac{1}{1 + e^{-(x)}}$$

Why not linear regression?

Predicted values may lie **outside** the interval [0, 1].

Assumption of **constant variance** is **violated** as variance depends on x through its influence on π.

- $var(\varepsilon) = \pi_i(x)[1 - \pi_i(x)]$
- Standard errors not valid, and conclusions from them misleading.

16.Websites or URLs

UNIT	CONTENT /TOPIC DETAILS	HYPERLINK DETAILS
UNIT-1	**R -Programming**	R: What is R? (r-project.org) R Programming | Coursera R Programming Fundamentals | edX
UNIT-2	**Basic Statistics**	Descriptive Statistics With R Software - Course (nptel.ac.in)
UNIT-3	**Predictive Analytics**	Data Science for Engineers - Course (nptel.ac.in)
UNIT-4	**Logistic Regression**	Machine Learning | Coursera
UNIT-5	**Decision Tree**	Introduction to Machine Learning (IITKGP) - Course (nptel.ac.in)