MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu.

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

	MUST KNOW CONCEPTS	**MKC**
DEPT - AI&DS		**2021-22**

Course Code & Course Name	**:**	**19ADCO5 / Introduction to Data Science**
Year/Sem/Sec	**:**	**II / III**

S.No.	**Term**	**Notation (Symbol)**	**Concept / Definition / Meaning / Units / Equation / Expression**	**Units**
Unit-I : Introduction				
1.	Data science		Data science involves gaining the knowledge from gathered data using different methods.	
2.	Data scientist		As a data scientist, you take a complex problem, research it, gather as a data, and we use to solve the problem.	
3.	Data Acquisition		It is a process of analysing the real world physical condition and converts into numerical values, which can be manipulated by computer.	
4.	Data preparation		Data preparation is a act of manipulating raw data into a form that can readily and accurately be analysed.	
5.	Data cleaning		Data cleaning is a process of identifying and correcting corrupt, incorrect and irrelevant data from reference set or table.	
6.	Data transformation		Data transformation is a process of converting data from one format or structure into another format or structure.	
7.	Handling outliers		Outliers are often used for the fraud detection and finding the malicious activities which happens on the field.	
8.	Data integration		In this, the data scientist ensures the data is accurate and reliable.	
9.	Data reduction		Data reduction is the transformation of numerical or alphabetical digital information derived experimentally into a corrected and simplified form.	
10.	Data mining		It is the process of extracting the required information from the larger set of raw data.	
11.	Model building		In this the process involves setting up ways of collecting data and finding a statistical, mathematical or a stimulation model to gain	

			understanding and make predications.	
12.	clustering		It is a task of dividing the population or data points into number of groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.	
13.	Essential data science skills		1. statistical analysis 2. machine learning 3. computer science and programming 4. data storytelling 5. business intuition 6. analytical thinking 7. critical thinking 8. inquistiveness 9. interpersonal skills	
14.	Statistical analysis		Identify patterns in data. This includes having a keen sense of pattern detection and normally detection.	
15.	Machine learning		Implement algorithms and statistical models to enable a computer automatically learn from data.	
16.	Computer science and programming		Applying the principle of AI, database system and software engineering. known to Write the programs like java, python and SQL programming languages.	
17.	Data storytelling		Data storytelling is the practice of building a narrative around a set of data and its accompanying visualizations.	
18.	Business intuition		Connect stakeholders to gain a full understanding of the problem they are looking to solve.	
19.	Analytical thinking		Find analytical solutions to abstract business issues.	
20.	Critical thinking		Apply objective analysis of facts before coming to a conclusion.	
21.	Inquistiveness		Look beyond what’s on the surface to discover patterns and solutions within the data	
22.	Interpersonal skills		Communicate across a diverse audience across all levels of an organization.	
23.	Fundamental steps to complete a data analytics project		Step 1: understand the business Step 2: get your data Step 3: explore and clean your data Step 4: enrich your dataset Step 5: build helpful visualizations Step 6: get predictive Step 7: iterate	
24.	Applications for data science		1. fraud and risk detection 2. healthcare 3. gaming 4. E-commerce 5. banking 6. transport	

			7. education	
25.	Jobs for data science		1. data scientists 2. data analyst 3. data engineering 4. business intelligence specialists 5. data architects	
Unit-II : Data Collection and Data Pre-Processing				
26.	Data collection		Data collection is the process of accumulating data that’s required to solve a problem statements.	
27.	Steps to collect the data		1. identify a problem statement 2. determine what data type is needed 3. decide on data sources 4. create a timeline 5. collect your data	
28.	Data pre-processing		Data preparation plays an important role in your workflow. You need to transform the data in a way that a computer would be able to work with it.	
29.	Steps in data pre-processing		1. data cleaning • Missing data • Noisy data 2. data transformation • Normalization • Attribute selection • Discretization • Concept hierarchy generation 3. data reduction • Data cube aggregation • Attribute subset selection • Numerosity reduction • Dimensionality reduction	
30.	Missing data		You may also notice that some important values are missing. These problems arise due to human factor, program errors and other reasons.	
31.	Noisy data		A large amount of additional meaningless data is called noisy data.	
32.	Normalization		Normalization is a technique often applied as part preparation for machine leaning. It is used while the features have different ranges.	
33.	Attribute selection		If you construct a new features combining the given features in order make the data mining process more efficient, it is called as attribute selection.	
34.	Discretization		Data discretization refers to a method of converting continuous data into discrete buckets by grouping it.	
35.	Concept hierarchy		Concept hierarchy generation based on the	

	generation		number of distinct values per attribute.	
36.	Aggregation		In the case of data aggregation, the data is pooled together and presented in a unified format for data analysis.	
37.	Numerosity reduction		Numerosity reduction is a method of data reduction that replaces the original data by a smaller form of data representation.	
38.	Types of numerosity reduction		1. parametric 2. non- parametric	
39.	Dimensionality reduction		Dimensionality reduction is the transformation of data from a high-dimensional space into low-dimensional space.	
40.	Data cleansing		Data cleaning is a process of identifying and correcting corrupt, incorrect and irrelevant data from reference set or table.	
41.	Steps involved for data cleansing		1. removal of unwanted observations 2. fixing structural errors 3. managing unwanted outliers 4. handling missing data	
42.	Tools for data cleansing		• Openrefine • Trifecta wrangler • TIBCO clarity • Cloudingo • IBM infosphere quality stage	
43.	Components of data integration		➢ Data migration ➢ Enterprise application integration (EAI) ➢ Master data management ➢ Data aggregation	
44.	Types of data aggregation		❖ data federation ❖ data warehousing	
45.	Data federation		Data is combined into virtual database.	
46.	Data warehousing		Data is combined into a physical database.	
47.	Advantages of data warehousing		1. improved business intelligence 2. rapid access to data 3. historical intelligence	
48.	Disadvantages of data warehousing		1. cost of scaling 2. maintenance cost	
49.	Challenges associated with MDM strategy		1. complexity 2. overlap 3. governance 4. standards	
50.	Categories of data integration		1. analytical data integration (AnDI) 2. operational data integration (OnDI) 3. hybrid data integration (HyDI)	
Unit III- Exploratory Data Analytics				
51.	Descriptive statistics		A population is the group to be studied, and population data is a collection of all in the	

			population.	
52.	Descriptive measures		Descriptive measures of populationare called parameters and typically using greek letters. The population mean is μ (mu).	μ
53.	Mean		The arithmetic mean of a variable, often called as average, is computed by adding up all the values and dividing by the total numbers of values.	
54.	Medium		The median of a variable is the middle of the data set when the data are sorted in order form least to greatest.	
55.	Mode		The mode is the value that appears frequently in the data set.	
56.	Range		The range is the difference between the highest and lowest values in a set of numbers.	
57.	Variance		The variance is the average of the squared differences from the mean.	
58.	Standard deviation		In statistics, the standard deviation is a measure of the amount of variance or dispersion of set of values.	
59.	Central limit theorum		In this theorem, the regardless of the shape of our population, the sampling distribution of the sample mean will be normal as the sample size increases.	
60.	Coefficient of variation		The coefficient of variation (CV) is a measure of relative variability**.** It is the ratio of the standard deviation to the mean.	
61.	Variability		Variability refers to how spread out; that is, it refers to the amount of spread of the scores around the mean.	
62.	Graphical representation		A graph is defined as a chart with statistical data, which represented in the form of curves or lines drawn across the coordinate point plotted on the surface.	
63.	Types of graphical representation		1. line graphs 2. bar graphs 3. histograms 4. line plots 4. frequency table etc…..	
64.	Advantages of graphical representation		• It makes data more easily understandable • It saves time • It makes the comparison of data more efficient.	
65.	Pie charts		A pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion.	
66.	Bar charts		A bar chart is a chart which represent the data in the rectangular box in the vertical position.	

67.	Histograms		A histogram is a bar graph like representation of data that buckets a range of outcomes into columns along the x-axis.	
68.	Skewness		Skewness is a measure of the symmetry of a distribution.	
69.	Types of skewness		1. positive skewed or right-skewed 2.negative skewed or left-skewed	
70.	Kurtosis		Kurtosis refers to the degree of presence of outliers in the distribution.	
71.	Excess kurtosis		The excess kurtosis is used in statistics and probability theory to compare the kurtosis is coefficient with that normal distribution. It can be positive, negative or near to zero.	
72.	Types of kurtosis		1. lepokurtic 2. platykurtic 3. mesokurtic	
73.	Pivot tables		Pivot table are a technique in data processing. They arrange and rearrange statistics in order to draw attention to useful information.	
74.	Two ways of ANOVA		1. two way ANOVA with replication 2. two way ANOVA without replication	
75.	Assumptions for two way ANOVA		• The population must be close to a normal distribution • Samples must be independent • Population variances must be equal • Groups must have equal sample sizes.	
UNIT -IV Model Development				
76.	Regression		It estimates the relationship between variables	
77.	Types of linear regression		• Simple linear regression • Multiple linear regression	
78.	Error function		It is the distance between current state and ideal state	
79.	Mean Squared Error	MSE = RSS / n	It is the mean of squared residuals and is calculated by dividing RSS by the number of data values	
80.	Root Mean Squared Error		It is the square root of mean squared error and is more suitable when large errors are particularly undesirable.	
81.	Mean Absolute Error		It is the measure of errors between paired observations expressing the same phenomenon	
82.	Ordinary Least Squares		It is a method in linear regression for estimating the unknown parameters by creating a model	
83.	Feature Selection		Certain features from the dataset are selected as the data is huge and multi dimensional used to better understand the data	
84.	Multi collinearity		It is a phenomenon in which one feature variable in a regression model is highly linearly correlated with another feature variable	

85.	Null Hypothesis		It is a type of hypothesis used in a statistics that proposes that there is no difference between certain characteristics of a population	
86.	Forward selection		It is a iterative method in which we start with having no feature in the model	
87.	Backward selection		It is a feature selection technique while building a machine learning model	
88.	Representation Learning		It is an area of research that focuses on how to learn compact , numerical representations for different sources of signal	
89.	Data Visualization		It is the process of translating large data sets and metricsninto charts , graphs and other visuals	
90.	Data Splitting		It is the acts of partitioning available data into two portions , usually for cross-validatory purposes	
91.	Data splitting purpose		There are two portions.One portion is used to develop a Predictive model and another portion is to evaluate the model's performance	
92.	Benefits of Data Visualization		• Increases the speed of decision making • Solves data inefficiencies an absorb vast amounts of data presented in visual formats • Identifies errors and inaccuracies in data quickly • Promotes storytelling and Conveys the right message to the audience • Optimize and instantly retrieve data via tailor-made reports • Explore business insights and achieve business goals	
93.	Data Science Process Flow		Line Chart , Histogram , piechart , Area plot , Scatter Plots , Hexbins Plot , Heat map , Box plot , Pair Plot , Bar Chart	
94.	Histogram		It is a graphical representation that organizes a group of data points into specified ranges	
95.	Characteristics of a Histogram		• Used to display Continuous data in a categorical form • No gaps between the bars , Unlike a bar graph • Width of the bins is equal	
96.	Linear Regression		It is a linear approach fo rmodelling the relationship between a scalar response and one or more explanatory Variables	
97.	Area Under Curve		It is a measure of the ability of a classifier to distinguish between classes and is used as a summary of ROC curve	
98.	Sensitivity		It is a metric that evaluates a model's ability to predict true positives of each available category	
99.	Specificity		It is metric that evaluates a model's ability to predict true negatives of each available category	
100.	Precision		It indicates the rate at which positive predictions	

			are correct	
UNIT-V Model Evaluation				
101.	Model Evaluation		Model Evaluation is the Subsidiary part of the model development process. It is the phase that is decided whether the model performs better.	
102.	Generalization		It refers to your model's ability to adapt properly to new , previously unseen data	
103.	Bias		Bias is the average squared difference between prediction and true values. It measure how good your model fits the data	
104.	Variance		If you train your data on training data and obtain a very low error , upon changing the data and then training the same previous model , you experience a high error , this is variance.	
105.	Regularization		It is a method to avoid high variance and overfitting as well as to increase generalization	
106.	Confusion Matrix		Confusion matrix is an N x N matrix , where N represents the number of categories in the target variable	
107.	Cost Of Classification		Cost of Classification is a measure of computing cost for classification models	
108.	Accuracy		It is the ratio of correct predicted values over the total predicted values	
109.	True Positive rate		$TPR = \frac{TP}{TP + FN}$	
110.	False Negative Rate		$FNR = \frac{FN}{TP + FN}$	
111.	True Negative Rate		$TNR = \frac{TN}{FP + TN}$	
112.	False Positive Rate		$FPR = \frac{FP}{FP + TN}$	
113.	Precision		$Precision = \frac{TP}{FP + TP}$ It is an evaluation metric which tells us out of all positive predictions , how many are actually positive	
114.	Recall		$Recall = \frac{TP}{FN + TP}$	
115.	F1 Score		$F1 = \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}}$ F1 is the harmonic mean of precision and recall	

116.	Log Loss		Log Loss is the negative average of the log of corrected-predicted probabilities for each instance	
117.	AUC-ROC		Area Under the Curve - Receiver Operating characteristics is an evaluation metric for binary classification which gives trade-off between false positive rate and true positive rate	
118.	Overfitting		Refers to a model can’t generalize or fit well on unseen data set.	
119.	Underfitting		Refers to a model that can neither model the training dataset nor generalize to new dataset.	
120.	Ridge Regression		It is a model tuning method that is used to analyze any data that suffers from multicollinearity	
121.	To Prevent Overfitting		• You need to add regularization in case of Linear and SVM models. • In decision tree models you can reduce the maximum depth. • While in Neural Networks, you can introduce dropout layer to reduce overfitting	
122.	To Prevent Underfitting		• Increase model complexity • Increase the number of features , performing feature engineering • Remove noise from the data • Increase the number of epochs or increase the duration of training to get better results	
123.	Logistic Regression		It is a supervised machine learning algorithm used to predict a dependent	
124.	Hyperparameter		These are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning	
125.	Parameter		It is a function argument that could have one of range of values	
Placement Questions				
126.	Three times the first of three consecutive odd integers is 3 more than twice the third. The third integer is:		Let the three integers be x, $x + 2$ and $x + 4$. Then, $3x = 2(x + 4) + 3 \Leftrightarrow x = 11$. $\therefore$ Third integer = $x + 4 = 15$.	
127.	Look at this series: 7, 10, 8, 11, 9, 12, ...		This is a simple alternating addition and subtraction series. In the first pattern, 3 is added; in the second, 2 is subtracted.	
128.	Look at this series: 22, 21, 23, 22, 24, 23, ….		In this simple alternating subtraction and addition series; 1 is subtracted, then 2 is added, and so on.	
129.	Look at this series: 53, 53, 40, 40, 27, 27, ...		In this series, each number is repeated, then 13 is subtracted to arrive at the next number.	

130.	Look at this series: 1.5, 2.3, 3.1, 3.9, ...		In this simple addition series, each number increases by 0.8.	
131.	Three times the first of three consecutive odd integers is 3 more than twice the third. The third integer is:		Let the three integers be x, $x + 2$ and $x + 4$. Then, $3x = 2(x + 4) + 3 \quad \Leftrightarrow \quad x = 11$. ∴ Third integer = $x + 4 = 15$.	
132.	Look at this series: 7, 10, 8, 11, 9, 12, ...		This is a simple alternating addition and subtraction series. In the first pattern, 3 is added; in the second, 2 is subtracted.	
133.	Look at this series: 22, 21, 23, 22, 24, 23, ...		In this simple alternating subtraction and addition series; 1 is subtracted, then 2 is added, and so on.	
134.	(112 x 5^4) = ?		(112 x 5^4) = 112 x(10)4=112 x 10^4=1120000= $7000022^4$16	
135.	It was Sunday on Jan 1, 2006. The day of the week Jan 1, 2010 is		On 31st December, 2005 it was Saturday. Number of odd days from the year 2006 to the year 2009 = (1 + 1 + 2 + 1) = 5 days. ∴ On 31st December 2009, it was Thursday. Thus, on 1st Jan, 2010 it is Friday.	
136.	Today is Monday. After 61 days, it will be:		Each day of the week is repeated after 7 days. So, after 63 days, it will be Monday. ∴ After 61 days, it will be Saturday.	
137.	If 6th March, 2005 is Monday,The day of the week on 6th March, 2004 is		The year 2004 is a leap year. So, it has 2 odd days. But, Feb 2004 not included because we are calculating from March 2004 to March 2005. So it has 1 odd day only. ∴ The day on 6th March, 2005 will be 1 day beyond the day on 6th March, 2004. Given that, 6th March, 2005 is Monday. ∴ 6th March, 2004 is Sunday (1 day before to 6th March, 2005).	
138.	The days in$x$ weeks x days?		x weeks x days = $(7x + x)$ days = $8x$ days.	
139.	On 8th Feb, 2005 it was Tuesday. The day of the week on 8th Feb, 2004 is		The year 2004 is a leap year. It has 2 odd days. ∴ The day on 8th Feb, 2004 is 2 days before the day on 8th Feb, 2005. Hence, this day is Sunday.	
140.	The greatest number that will divide 43, 91 and 183 so as to leave the same remainder in each case.		Required number = H.C.F. of (91 - 43), (183 - 91) and (183 - 43) = H.C.F. of 48, 92 and 140 = 4.	
141.	The H.C.F. of two numbers is 23 and the other two factors of their L.C.M. are 13 and 14. The larger of the two numbers		Clearly, the numbers are (23 x 13) and (23 x 14). ∴ Larger number = (23 x 14) = 322	

	is:			
142.	$(112 \text{ x } 5^4) = ?$		$(112 \text{ x } 5^4) = 112 \text{ x}(10)4=112 \text{ x } 10^4=1120000= 7000022^{4}16$	
143.	It was Sunday on Jan 1, 2006.The day of the week Jan 1, 2010 is		On 31st December, 2005 it was Saturday. Number of odd days from the year 2006 to the year 2009 = (1 + 1 + 2 + 1) = 5 days. ∴ On 31st December 2009, it was Thursday. Thus, on 1st Jan, 2010 it is Friday.	
144.	Today is Monday. After 61 days, it will be:		Each day of the week is repeated after 7 days. So, after 63 days, it will be Monday. ∴ After 61 days, it will be Saturday.	
145.	If 6th March, 2005 is Monday, The day of the week on 6th March, 2004 is		The year 2004 is a leap year. So, it has 2 odd days. But, Feb 2004 not included because we are calculating from March 2004 to March 2005. So it has 1 odd day only. ∴ The day on 6th March, 2005 will be 1 day beyond the day on 6th March, 2004. Given that, 6th March, 2005 is Monday. ∴ 6th March, 2004 is Sunday (1 day before to 6th March, 2005).	
146.	The sin x weeks x days?		x weeks x days = $(7x + x)$ days = $8x$ days.	
147.	On 8th Feb, 2005 it was Tuesday. The day of the week on 8th Feb, 2004 is		The year 2004 is a leap year. It has 2 odd days.∴ The day on 8th Feb, 2004 is 2 days before the day on 8th Feb, 2005. Hence, this day is Sunday.	
148.	Find the greatest number that will divide 43, 91 and 183 so as to leave the same remainder in each case.		Required number = H.C.F. of (91 - 43), (183 - 91) and (183 - 43) = H.C.F. of 48, 92 and 140 = 4.	
149.	The H.C.F. of two numbers is 23 and the other two factors of their L.C.M. are 13 and 14. The larger of the two numbers is:		Clearly, the numbers are (23 x 13) and (23 x 14). ∴ Larger number = (23 x 14) = 322	
150.	Two trains running in opposite directions cross a man standing on the platform in 27 seconds and 17 seconds respectively and they cross each other in 23 seconds. The ratio of their speeds is:		Let the speeds of the two trains be x m/sec and y m/sec respectively. Then, length of the first train = $27x$ meters, and length of the second train = $17y$ meters. $\therefore \frac{27x + 17y}{x+ y} = 23$ $\Rightarrow 27x + 17y = 23x + 23y$ $\Rightarrow 4x = 6y$ $\Rightarrow \frac{x}{y} = \frac{3}{2}$.	

Faculty Team Prepared	Signatures
1. Dr.P.Srinivasan	

HoD