Gandhinagar Institute of Technology

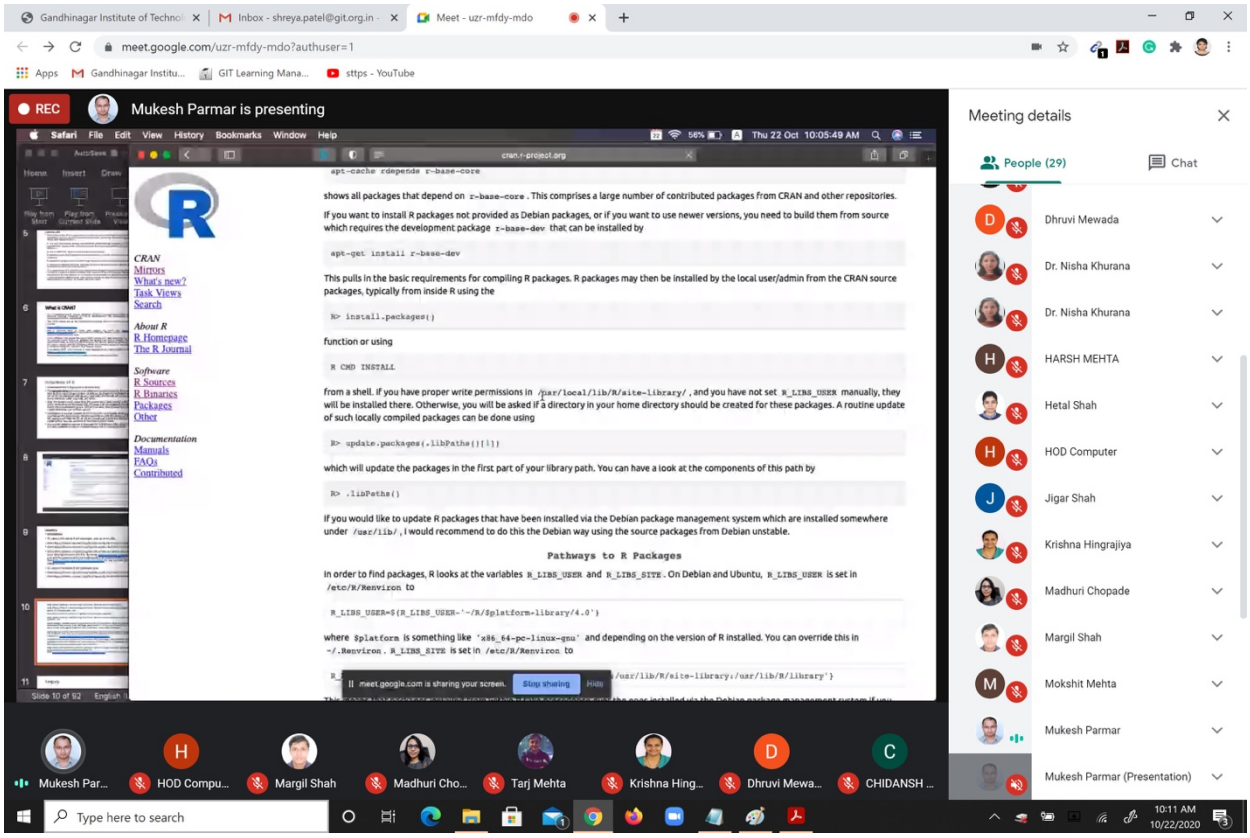
A Report on
"Hands on Machine Learning Using R"
Sponsored by GUJCOST and DST
(From 22nd October to 23rd October 2020)

Objective:
The objective of the webinar was to provide the platform for applying various computer science concepts such as Machine Learning and Artificial Intelligence using R. The goal of the workshop was to empower the participants for developing various things using Machine Learning. Participants would get the knowledge of R and its use in Statistical and Machine Learning applications with the help of hands-on exercises.

About Webinar:
The Webinar provided a valued platform for the industry specialists and academicians from institutes to exchange and explore the latest views on technological developments in the field of Machine Learning. 107 participants have been registered from different institutes. These technologies have been buzz word now a days but the main challenge is to understand the actual working and implementation in each of them. This Workshop not only provided theoretical knowledge regarding subject but it also provided hands on sessions in each area.

Day-1(22/10/20) Time (09:00 am to 03:45 pm)
Webinar started with Dr.Rajan Patel, Associate Professor of Gandhinagar Institute of Inaugural Speech. Dr. Rajan Patel discussed about data as one of the important features of every organization because it helps business leaders to make decisions based on facts, statistical numbers and trends. Due to this growing scope of data, data science came into picture which is a multidisciplinary field. It uses scientific approaches, procedure, algorithms, and framework to extract the knowledge and insight from a huge amount of data. The extracted data can be either structured or unstructured. Data science is a concept to bring together ideas, data examination, Machine Learning, and their related strategies to comprehend and dissect genuine phenomena with data. Data science is an extension of various data analysis fields such as data mining, statistics, predictive analysis and many more. Data Science is a huge field that uses a lot of methods and concepts which belongs to other fields like information science, statistics, mathematics, and computer science. Some of the techniques utilized in Data Science encompasses machine learning, visualization, pattern recognition, probability model, data engineering, signal processing, etc.

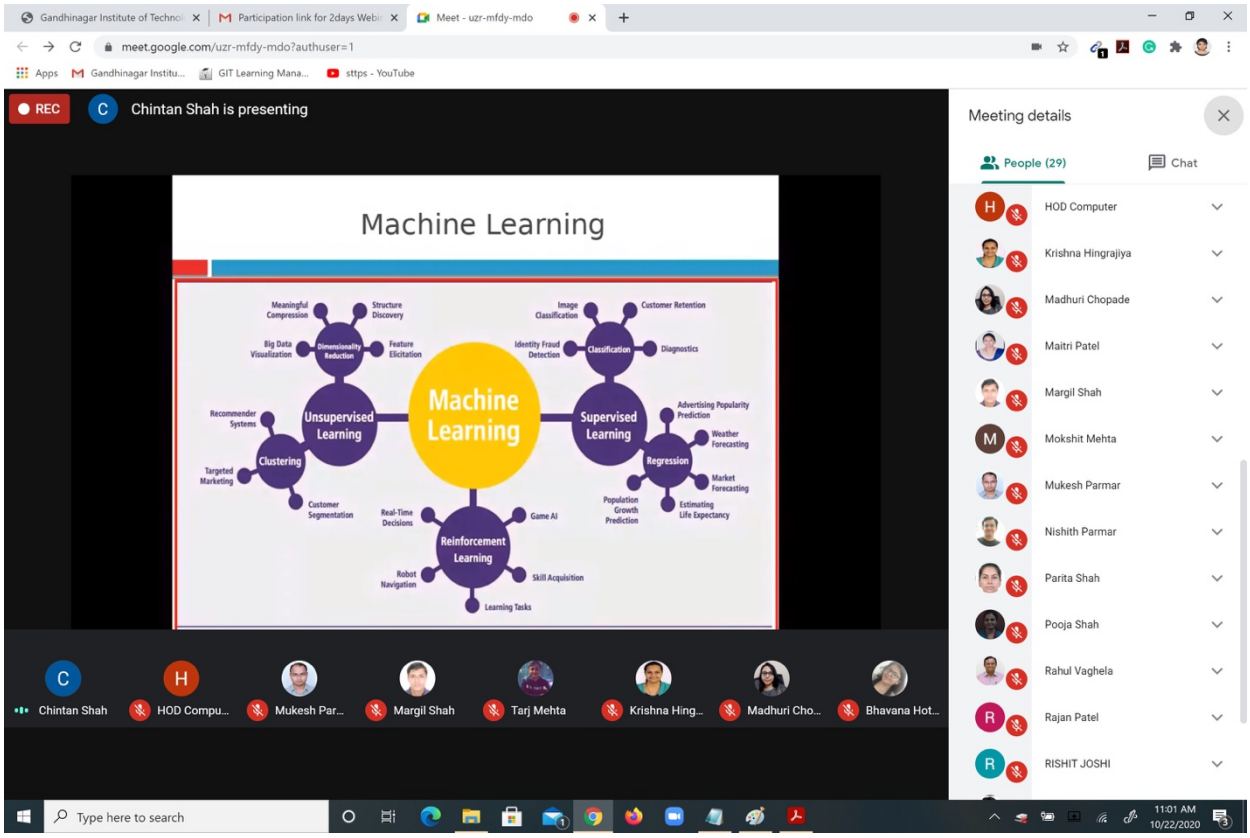
Speaker 1: Prof Mukesh Parmar (Asst. Professor, Gandhinagar Institute of Technology)
Time: 9:30 am to 10:30 am
Prof. Mukesh Parmar is working as Assistant Professor, Computer Engineering Department, Gandhinagar Institute of Technology. He has more than 15years of experience in IT industry as well as academia. He has expertise in the area of Image Processing, Machine learning and data science.



Prof. Mukesh Parmar discussed about R Introduction and Installation. R' is a programming language for data analysis and statistics. It is free, and very widely used by professional statisticians. It is also very popular in certain application areas, including bioinformatics. R is a dynamically typed interpreted language and is typically used interactively. It has many built-in functions and libraries, and is extensible, allowing users to define their own functions and procedures using R, C or Fortran. It also has a simple object system. He gave presentation on installation on Windows, Mac and different Linux based operating system.

Speaker 2: Dr. Chintan Shah (Data Scientist, Corpository, Ahmedabad)
Time: 10:45 am to 12:45 pm

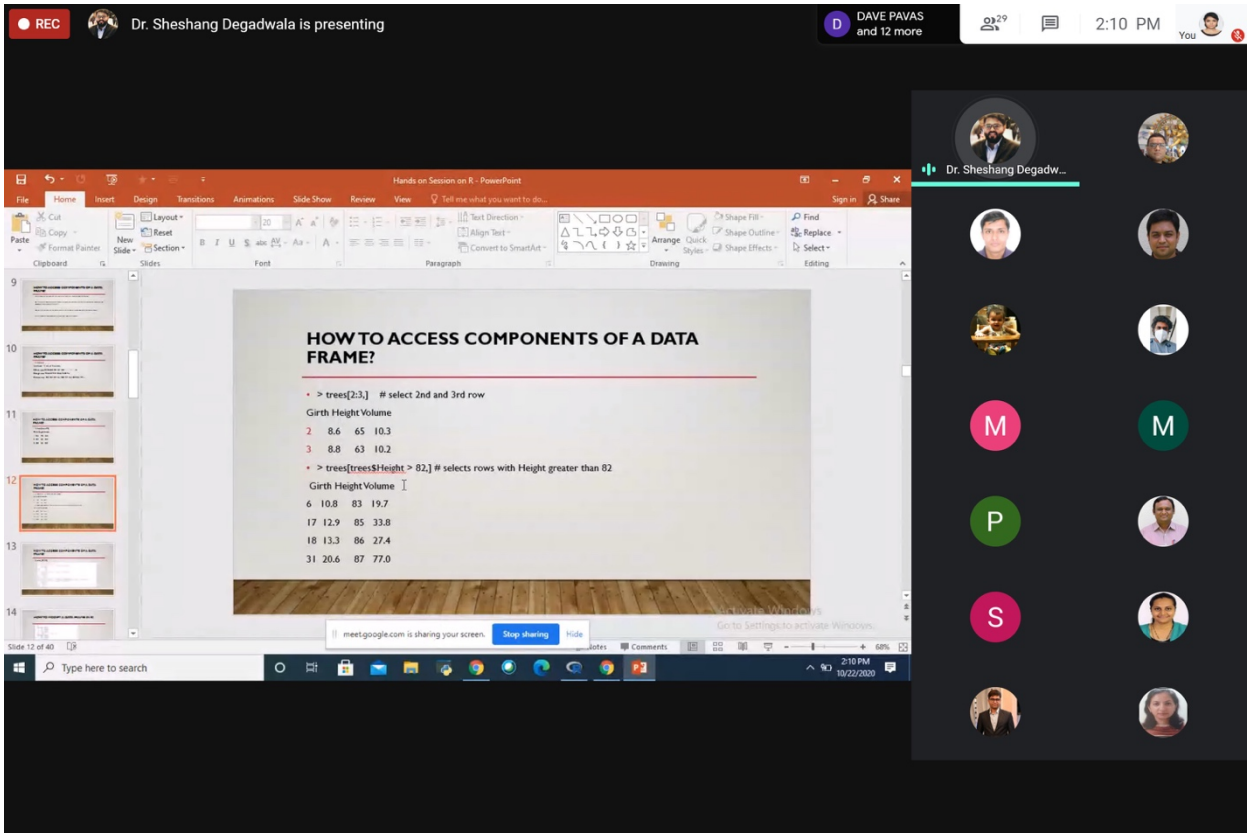
Dr. Chintan Shah is Data Scientist at Corpository, Ahmedabad. He has expertise in information retrieval using BM25 score technologies and NLP. His focused area is to provide solutions to real world problems using cutting edge AI/ML/Data Science/NLP/BI technologies. He has completed his PhD from MS University, Baroda. He is also having more than 9 years of experience in academia. His research areas of interest are Machine Learning, Text Summarization with Semantic Analysis.



Dr. Chintan Shah discussed about Introduction to Machine Learning, Supervised Classification (Naïve Bayes, Decision Trees, Random Forest and SVM) Classification. Supervised learning is where you have input variables (x) and an output variable (Y), and you use an algorithm to learn the mapping function from the input to the output. It is called supervised learning because the process of an algorithm learning from the training dataset can be thought of as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes predictions on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

Speaker 3: Dr. Sheshang Degadawala Associate Professor and Head of Computer Engineering Department, Sigma Institute of Engineering, Vadodara
Time: 1:45 pm to 3:45 pm

Dr. Sheshang D. Degadwala is presently working as Associate Professor and Head of Computer Engineering Department, Sigma Institute of Engineering, Vadodara. He obtained his B.E. Degree from the department of Computer Engineering, BITs, Vadodara. Subsequently, He obtained his MTech. Degree From Charusat University, Changa and Completed his Ph.D. in Computer Engineering from Madhav University, Abu Road, Sirohi, Rajasthan, India in year 2018. He has published 112 research papers in reputed international journals and conferences including IEEE, Elsevier and Springer. His main research work focuses on Image Processing, Computer Vision, Information Security, Theory of Computation and Data Mining. He is also Microsoft Certified in Python Programming. He has published 8 books and he got grant for 1 patent. He has received 35 awards for academic and research achievement.

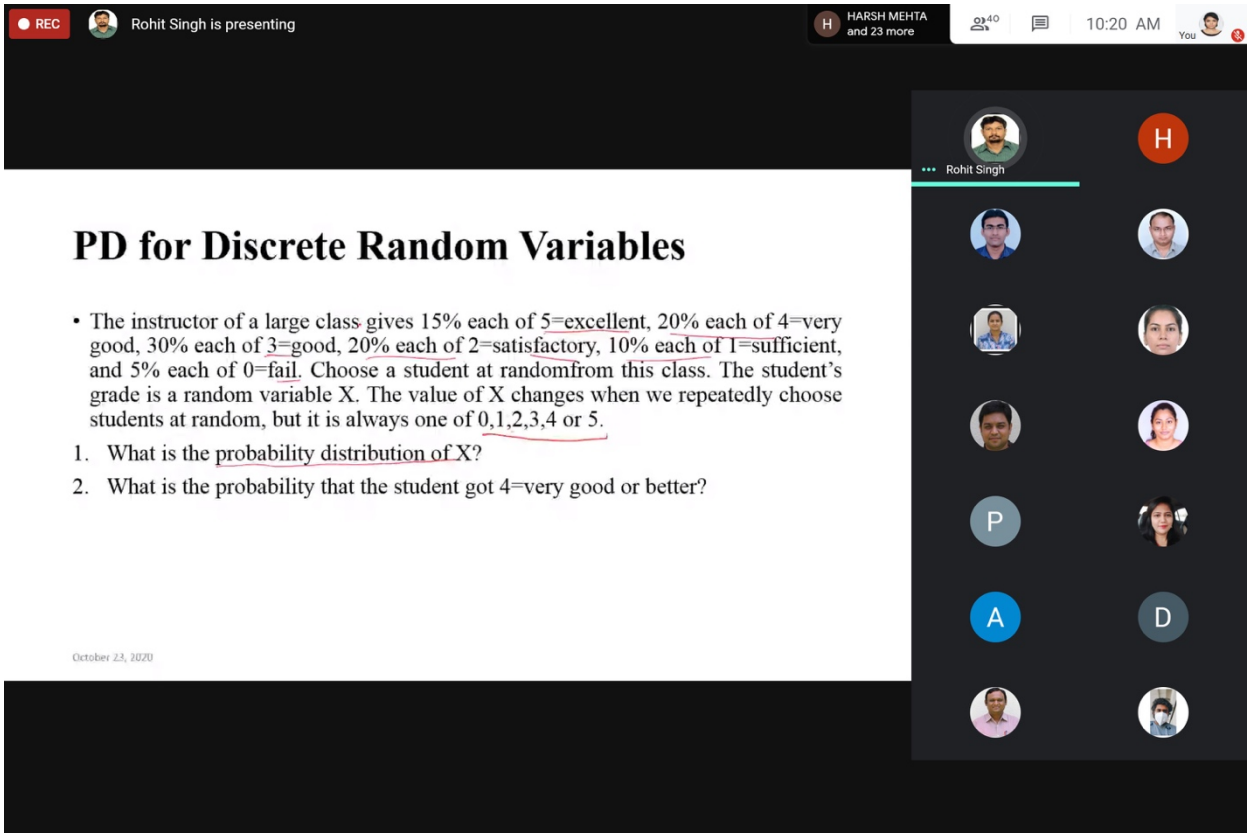


Dr. Sheshang D. Degadwala discussed about RStudio and be able to read in data and write out data files from various sources, basic data structures, analysing the data, Sampling the data, plotting the with different types of plot. Importance of sampling is to make a random sample of a set according to a probability distribution among the elements of the set. In the case of a training batch, we will attach weights to the training examples, and the elements of high weight will have more chance to be selected. Data visualization is the representation of data or information in a graph, chart, or other visual format. It communicates relationships of the data with images. This is important because it allows trends and patterns to be more easily seen. With the rise of big data upon us, we need to be able to interpret increasingly larger batches of data. Machine learning makes it easier to conduct analyses such as predictive analysis, which can then serve as helpful visualizations to present. But data visualization is not only important for data scientists and data analysts, it is necessary to understand data visualization in any career. Whether you work in finance, marketing, tech, design, or anything else, you need to visualize data. That fact showcases the importance of data visualization.

Day-2(23/10/20) Time (09:30 am to 3:45 pm)

Speaker 1: Prof Rohit Singh (Asst. Professor, Gandhinagar Institute of Technology)
Time: 9:30 am to 10:30 am

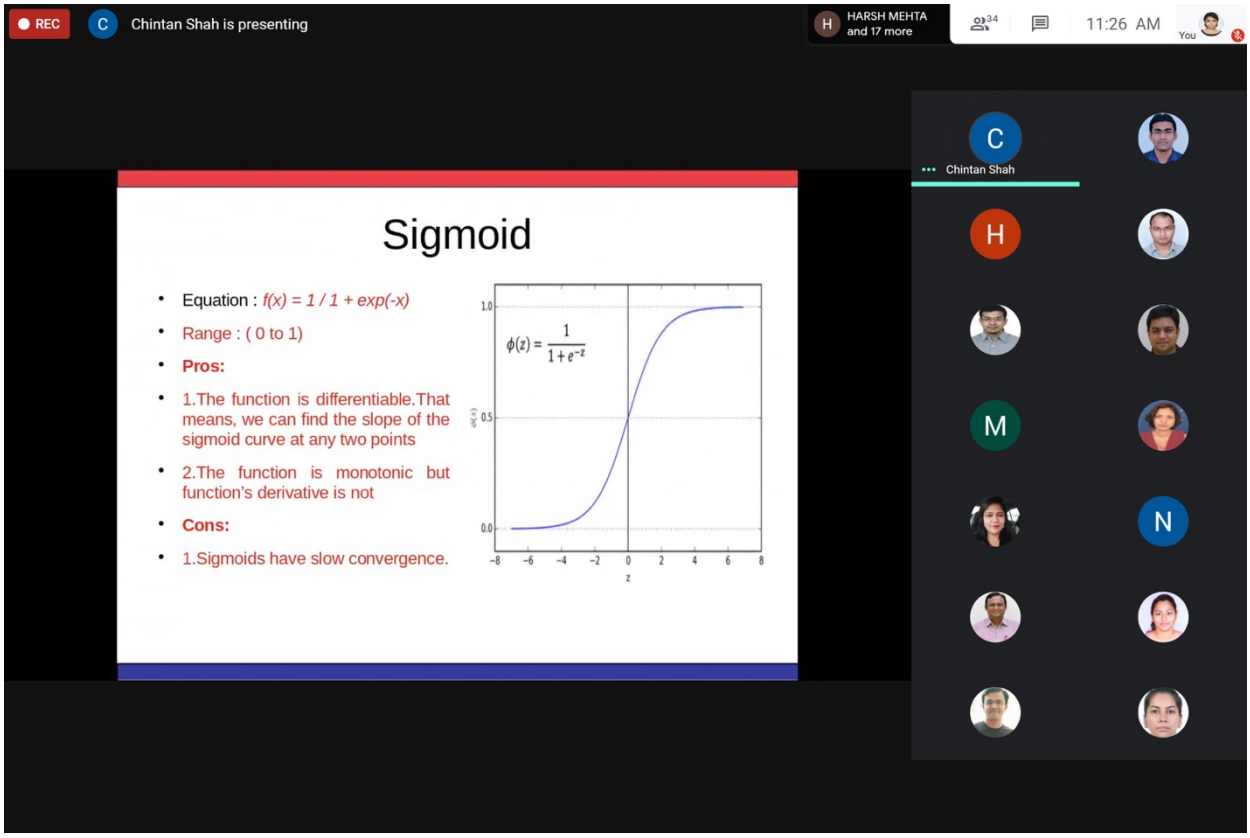
Prof. Rohit Singh is working as Assistant Professor, Computer Engineering Department, Gandhinagar Institute of Technology. He has 10 years of experience in the academia. His research area of interest is machine learning and data science.



Prof. Rohit Singh discussed about Basic Statistics, Probability Distribution and Linear Algebra. Linear Algebra is a branch of mathematics that lets you concisely describe coordinates and interactions of planes in higher dimensions and perform operations on them. Think of it as an extension of algebra (dealing with unknowns) into an arbitrary number of dimensions. Linear Algebra is about working on linear systems of equations (linear regression is an example: y = Ax). Rather than working with scalars, we start working with matrices and vectors (vectors are really just a special type of matrix). Uncertainty involves making decisions with incomplete information, and this is the way we generally operate in the world. Handling uncertainty is typically described using everyday words like chance, luck, and risk. Probability is a field of mathematics that gives us the language and tools to quantify the uncertainty of events and reason in a principled manner. It would be fair to say that probability is required to effectively work through a machine learning predictive modelling project. Machine learning is about developing predictive models from uncertain data. Uncertainty means working with imperfect or incomplete information.

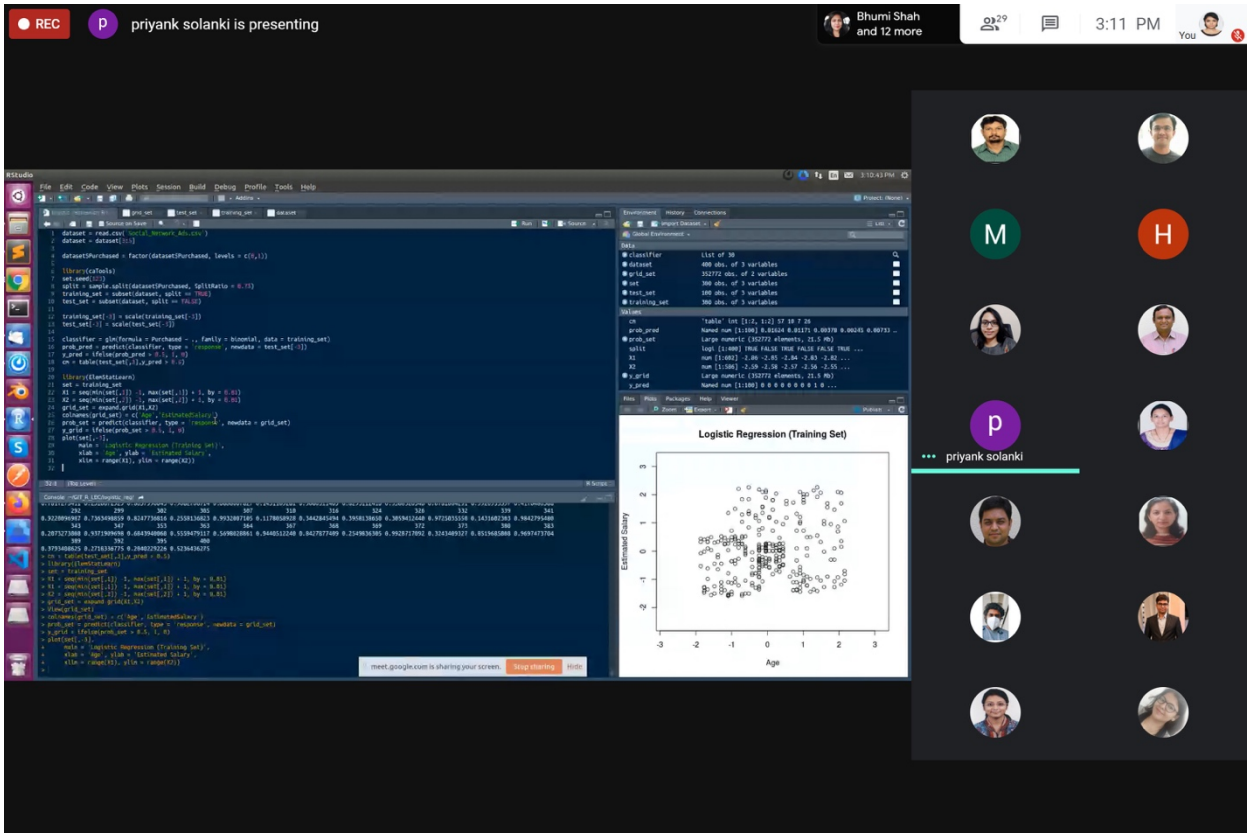
Speaker 2: Prof Chintan Shah (Data Scientist, Corpository , Ahmedabad)
Time: 10:45 am to 12:45 pm

Dr. Chintan Shah is Data Scientist at Corpository, Ahmedabad. He has expertise in information retrieval using BM25 score technologies and NLP. His focused area is to provide solutions to real world problems using cutting edge AI/ML/Data Science/NLP/BI technologies. He has completed his PhD from MS University, Baroda. He is also having more than 9 years of experience in academia. His research areas of interest are Machine Learning, Text Summarization with Semantic Analysis.



Dr. Chintan Shah discussed about ANN (Artificial Neural Network) and CNN (Convolution Neural Network). Artificial Neural Network (ANN) uses the processing of the brain as a basis to develop algorithms that can be used to model complex patterns and prediction problems. In our brain, there are billions of cells called neurons, which processes information in the form of electric signals. External information/stimuli is received by the dendrites of the neuron, processed in the neuron cell body, converted to an output and passed through the Axon to the next neuron. The next neuron can choose to either accept it or reject it depending on the strength of the signal. The network architecture has an input layer, hidden layer (there can be more than 1) and the output layer. It is also called MLP (Multi-Layer Perceptron) because of the multiple layers. The hidden layer can be seen as a "distillation layer" that distills some of the important patterns from the inputs and passes it onto the next layer to see. It makes the network faster and efficient by identifying only the important information from the inputs leaving out the redundant information. Convolution layers are used to extract the features from input training samples. Each convolution layer has a set of filters that helps in feature extraction. In general, as the depth of CNN model increases, complexity of features learnt by convolution layers increases. For example, first convolution layer captures simple features while the last convolution layer captures complex features of training samples.

Speaker 3: Mr. Priyank Solanki (Webline India, AI/ML Engineer)
Time: 1:45 pm to 3:45 pm
Mr. Priyank Solanki is working as AI/ML Engineer in Webline India. He is providing expertise in data storage structures, data mining, and data cleansing, translating numbers and facts to inform strategic business decisions, Analysing sales figures, market research, logistics, or transport data.



Mr. Priyank Solanki implemented Linear Regression and Logistic Regression in hands on session of R. Linear regression and logistic regression both are machine learning algorithms that are part of supervised learning models. Since both are part of a supervised model so they make use of labelled data for making predictions. Linear regression is used for regression or to predict continuous values whereas logistic regression can be used both in classification and regression problems, but it is widely used as classification algorithm. Regression models aim to project value based on independent features. The main difference that makes both different from each other is when the dependent variables are binary logistic regression is considered and when dependent variables are continuous then linear regression is used. Every person must have come across linear models when they were at school. Mathematics taught us about linear models. It is the same model that is used widely in predictive analysis now. It majorly tells about the relationship between a target that is dependent variable and predictors using a straight line. Linear regression is basically of two types that are Simple and Multiple Regression.