Course Overview-7

**Python For Data Science**

In this, we introduce you to python basics required for data science. Libraries like bumpy pandas and matplotlib are discussed in detail.

The use of various kinds of plots for data visualization and their effective usage are also discussed.

**Exploratory Data Analysis and Data Preprocessing**

* Exploratory Data Analysis (EDA)\_Outline
* Data, Datatypes and Variables
* Central Tendency and Dispersion
* 5 point summary and skewness of data
* Box-plot, covariance and Coeff of Correlation
* Let's get our hands dirty with code
* Univariate and Multivariate Analysis
* Encoding Categorical Data
* Scaling and Normalization
* What is Preprocessing?
* Imputing missing values
* Working with Outliers
* Case study Analysis

**Intro to supervised learning**

Get introduced to Supervised machine learning. First, we get a sense of terms like, Model, train set, test set, features, and feature space, then we take a deeper look into the process of building and assessing a supervised machine learning model in a regression setting. And finally, we dive into one particular algorithm in supervised machine learning – Linear Regression, along with the advantages and limitations of using a linear regression model.

Concepts of machine learning:

* Definition of machine learning.
* Examples of use cases of machine learning.
* ML prerequisites.

Mathematical/feature space:

* Expressing data as points in a mathematical space.
* Defining a “model” which is always expressed as a mathematical equation.

Supervised ML intro:

* Characteristics of Supervised Machine Learning
* Steps to go through to build a reliable model

Linear regression and Pearson’s coefficient:

* Pearson’s coefficient equation and its
* Understanding R-value.

Linear regression – Mathematical concept

* Linear regression equation
* Best fit line
* Actuals, predicted and residuals
* SSE, SST and SSR. SSE + SSR = SST
* Coefficient of determination - R2

Linear regression advantages and disadvantages

**Applying linear regression on a dataset**

Overview: By the end of week 2 material, you will have a good idea on how to get started on a dataset. Starting with a high-level exploration of the data, we deal with missing values and perform univariate and bivariate analysis. Then we implement a linear regression model on a tabular dataset with a mix of categorical and continuous variables and then assess the model using certain metrics.

Descriptive analysis on the dataset

* Exploring the data at a high level
* Dropping variables that are most obvious of not being of any significant help for the model prediction.
* Creating dummy variables out of a categorical column after renaming the numerical labels with actual names.
* Issue of multicollinearity
* Consider the possibility of multiple Gaussians

Missing value imputation

* Recognizing columns that have missing values which aren’t necessarily always a blank space or a np.nan value.
* Filling the missing values (in our case with median)
* Data type check for all columns

Bivariate analysis

* Bivariate analysis of every variable against every other variable using a pairplot.
* Observing the univariate distribution of each variable.

Building a model using all information

* Separate the target column from the predictors and create individual DataFrames for both.
* Splitting data into train and test sets using the train\_test\_split function.
* Fitting a linear regression model to the training sets.
* Predict and assess the model using the test sets.
* Improving performance by accounting for interactions of the target column with other columns.
* Transforming input features using “PolynomialFeatures” from sci-kit learn.

Understanding data and relationship between variables

* Getting back to the basics of regression line using a toy dataset with a single predictor variable.

Error analysis and Adjusted R2

* Addressing R2 (Coefficient of determinant) at a deeper level
* Adjusted R2

Fluke correlation

* Fluke correlation
* Significance of adjusted R2

**Classification models – Logistic and Naïve Bayes**

Overview: Introduction to machine learning algorithms for a classification setting. A complete understanding of two of the most popular algorithms for a classification problem – Logistic regression and Naïve Byes’ classifiers. Confusion matrix and evaluation metrics in a classification setting. Hands-on exercises for performing classification using Logistic and Naïve Bayes’ classifiers.

Classification Algorithm – Logistic Regression

* A very brief discussion on probability
* Concept of odds
* Logistic regression equation expressed as log of odds of the probability of a certain class occurring
* High-level discussion on the workflow of logistic regression

Logistic regression model and Sigmoid function

* Logistic regression as an improvised version of linear regression for a classification setting.
* The necessity of using a sigmoid curve instead of a linear model.
* Log-loss function.
* Interpretation of the ‘S’ curve.
* Effect of misclassification and correct classification on the log-loss function.

Logistic regression, Confusion Matrix, Precision and Recall: Hands-on

* Exploration of Pima-Indian-diabetes dataset
* Building a logistic regression model after splitting the data into test and train sets
* Model evaluation using Confusion matrix metrics – Recall, Precision

Bayes’ Theorem (Later videos for a more detailed discussion)

* Naïve Bayes’ classifier and it’s a mathematical equation
* Advantages and limitations of Naïve Bayes’ classifier

Introduction to Naïve Bayes’ classifier

* Characteristics of NB algorithm
* Review of Conditional and joint probabilities

Naive Bayes' Classifier and examples

* Bayes’ theorem in detail with examples

Hands-on Exercise - Naive Bayes

* Hands-on exercise for Naïve Bayes algorithm on Pima-Indian dataset

**kNN (k Nearest Neighbors)**

Overview: Introduction to ‘ k- Nearest Neighbors ’ algorithm. Later we discuss variants of k-NN algorithm, Limitations and advantages of the algorithm,  way to determine the ‘k’ in kNN, and Various types of distance measurements. We demonstrate kNN on a dataset and later evaluate the performance of the algorithm using a confusion matrix.

Introduction to KNN and Euclidean distance

* Characteristics of the kNN algorithm
* Data as points within the mathematical space.
* Euclidean distance in detail

Determining Neighbors (k)

* A variant of kNN algorithm – Radius Neighbor Classifier
* Challenges in determining ‘k’

Find Nearest Neighbors – Approach

* kNN for regression
* The variant of kNN algorithm – KD tree based nearest neighbor
* Need for normalizing the data before applying kNN for classification or regression

Distance calculations and Pros/Cons of KNN

* Different ways of measuring distances
* Advantages and disadvantages of kNN

Code Walk-through – KNN

* Initiating kNN algorithm with some initial parameters
* Importing the dataset and comments on the high-level attributes of the dataset like shape, data types

EDA – KNN

* High-level exploration of the data in excel

Descriptive Statistics in Python – KNN

* A deeper exploration of data by computing the statistics
* Detecting outliers
* Comments on the class distribution
* The normalization of data after separating the target from the predictors

Splitting Function

* Using the train\_test\_split() function to split the data into test and train
* Fitting the kNN model to the train set
* Making predictions on the test set
* Building a confusion matrix out of the predicted and actual labels

Confusion Matrix using Seaborn

* Confusion matrix in detail
* Deep dive into the data using a pairplot

Data Cleaning - Model improvement

* Improving the model by accounting for the evidence we got from EDA until now

**SVM (Support Vector Machines)**

Overview: We discuss the origins of SVMs, as a solution to overcome the limitations of perception. Also, Visualize an SVM in the mathematical space. We discuss K-SVM as an improvised version of SVM to deal with data that is not linearly separable. We discuss various kernels available to perform kernel tricks to implement K-SVM.

Support Vector Machines – Introduction

* Origin of SVM
* Discussion on linearly separable data
* Mathematically compute the position(above or below) of a data point w.r.t certain linear function (say a line)

Perceptron algorithm - Line properties

* Mathematics of adjusting the decision boundary using the slope vector
* Limitation of the perception
* Need for SVM to overcome the limitations of the perceptron

SVMs in 2D Space

* How SVM finds the optimal slope by maximizing the distance of the line to the nearest data point
* Hard margin vs Soft margin SVMs

SVM Kernel and Mercer's theorem

* Need for Kernel-SVM
* Mercer's theorem
* KSVM at a high level

Optimal hyperplane concept

* Some more characteristics of SVM

SVM - Hands-on exercise

* Applying SVM on Letters.csv dataset where we separate handwritten letters of the alphabet based on certain attributes.
* Summary of SVM
* Kernel functions
* Dealing with non-linearly separable data

**Unsupervised Learning**

In unsupervised learning, we discuss K-Means clustering and the sage of it. The use of means in real-life scenarios, and different terms associated with it are also duly explained along with a hands-on exercise.

**Decision Trees**

**Topic Objectives**

After completing this course, you will be able to:

* Understand and apply the Decision Tree algorithm for an appropriate problem
* Use the key Ensemble techniques to ensure the best performing model with minimal risk. The objectives envisaged in this course will be met through Lecture videos, Case Studies, Practice Exercises, and Mentored Learning Sessions. You will be exposed to Python Programming for analysis and Model building.

**Course Content**

 Decision Tree

* Decision Tree Classifier
* Gini Index
* Pruning

**Data Visualization Using Tableau (DVT)**

**Course Description**

The fastest way for people to understand data of any size is through a visual medium. Unfortunately, a lot of companies adopt visualization as no more than snazzy graphics. Very often the people who design information displays are not trained to design them for effective communication.   
A primary goal of data visualization is to communicate information clearly and efficiently to users via information graphics. Effective visualization helps users in analyzing and reasoning about data and evidence.   
This course will help you in understanding effective visualization using Tableau.

**Course Objectives**

The course objectives are to help develop a working proficiency of statistical concepts for decision making:

* Understand how the brain perceives data and draw insights from data
* Recognize visual characteristics of data that are meaningful
* Choose an appropriate display mechanism to communicate data effectively
* Transform data into information that is actionable
* Start using Tableau software to solve real business problems

**Course Prerequisites**

A thorough understanding of basic and advanced statistical concepts taught in the core and foundations module

**Pedagogy**

The objectives envisaged in this course will be met through Video Lectures, Case Studies, and Computer Exercises. You will be exposed to Tableau Software for data analysis using visualizations.

**Course Content**

* The objectives envisaged in this course will be met purely through practical demonstrations.

**Know who gonna teach**

Know Your Faculty-6

Python for Data science- Bharani Akella is a data science and Artificial Intelligence enthusiast working for Great Learning. He has been actively involved in various course development activities associated with data science, machine learning, and artificial intelligence. He possesses unique skillsets when it comes to programming and has proficiency in Python, R, Java, and SQL. Bharani is equally proficient when talking about core concepts in machine learning and AI. He has a unique style of content delivery and is well-loved by GL students and alumni for his teaching.

Exploratory Data Analysis - Abhinanda Sarkar Dr. Abhinanda Sarkar is the Academic Director at Great Learning for Data Science and Machine Learning Programs. Dr. Sarkar received his B.Stat. and M.Stat. degrees from the Indian Statistical Institute (ISI) and a Ph.D. in Statistics from Stanford University. He has taught applied mathematics at the Massachusetts Institute of Technology (MIT); been on the research staff at IBM; led Quality, Engineering Development, and Analytics functions at General Electric (GE); served as Associate Dean at the MYRA School of Business; and co-founded OmiX Labs. Dr. Sarkar’s publications, patents, and technical leadership have been in applying probabilistic models, statistical data analysis, and machine learning to diverse areas such as experimental physics, computer vision, text mining, wireless networks, e-commerce, credit risk, retail finance, engineering reliability, renewable energy, and infectious diseases, His teaching has mostly been on statistical theory, methods, and algorithms; together with application topics such as financial modeling, quality management, and data mining. Dr. Sarkar is a certified Master Black Belt in Lean Six Sigma and Design for Six Sigma. He has been visiting faculty at Stanford and ISI and continues to teach at the Indian Institute of Management (IIM-Bangalore) and the Indian Institute of Science (IISc). Over the years, he has designed and conducted numerous corporate training sessions for technology and business professionals. He is a recipient of the ISI Alumni Association Medal, IBM Invention Achievement Awards, and the Radhakrishan Mentor Award from GE India.

Data Visualization with Tableau - Vivek Anand Vivek Anand is a data visualization consultant with 10 years of experience. Vivek has an MBA from Monash University Melbourne Vic. The area of specialization includes Marketing & Econometrics. He has worked as Sales & Marketing professional handling teams of leading Indian hospitality brands across the country. His most recent assignment was for India’s largest Luxury hotel by ITC hotels in Chennai. He is a qualified trainer of Tableau 9.0 and has a passion for teaching. Vivekanand Profile:(https://in.linkedin.com/in/vivek0205

Supervised and Unsupervised learning: Prof. Mukesh Rao - (Consultant, Big Data & Machine Learning) Prof. Mukesh Rao is an Adjunct Faculty at Great Lakes for Big Data and Machine Learning. Mukesh has over 20 years of industry experience in Market Research, Project Management, and Data Science. Mukesh has conducted over 100 corporate training. Data Science training covers all the stages of CRISP-DM, tools, and techniques used in each stage, machine learning algorithms, and their application. Big Data training covers core Apache Hadoop technologies including HDFS, YARN, Map Reduce, PIG, HIVE, SQOOP, FLUME, SPARK, and MongoDB

**PRE requeseted**

Dear Learner,

This course consists of Machine learning and visualization techniques, for which you will require to have prior knowledge on Python and Statistics.

We do have various courses in our Great Learning academy platform. You can go and explore from the below link.

<https://www.greatlearning.in/academy>

Recommended Links:

Python: <https://www.greatlearning.in/academy/learn-for-free/courses/introduction-to-programming>

Statistics: <https://www.greatlearning.in/academy/learn-for-free/courses/statistics-for-machine-learning>