CERTIFIED DATA SCIENTIST ROAD MAP

PROGRAM SYLLABUS

CONTENTS

COURSE 1	DATA SCIENCE FOUNDATION	
MODULE 1	INTRODUCTION TO DATASCIENCE	WHAT IS DATA SCIENCE?
		EVOLUTION OF DATA SCIENCE
		DATA SCIENCE TERMINOLOGIES
MODULE 2	DATA SCIENCE VS BUSINESS ANALYTICS VS BIG DATA	COMPARING VARIOUS RELATED DOMAINS WITH DATA SCIENCE
MODULE 3	CLASSIFICATION OF BUSINESS ANALYTICS	DESCRIPTIVE ANALYTICS
		PREDICTIVE ANALYTICS
		DISCOVERY ANALYTICS AND PRESCRIPTIVE ANALYTICS
MODULE 4	DATA SCIENCE PROJECTWORKFLOW	CRIPS – DM FRAMEWORK
		DATA SCIENCE PROJECT WORKFLOW
MODULE 5	ROLES IN DATA SCIENCE	INDUSTRY ROLES AND RESPONSIBILITIES
MODULE 6	APPLICATION OF DATASCIENCE IN VARIOUS INDUSTRIES	INDUSTRY ADOPTION: HEALTH CARE, FINANCE & BANKING, MANUFACTURING, RETAIL, LOGISTICS,HUMAN RESOURCE
		KEY USE CASES
		TRENDS ON DATA SCIENCE ADOPTION

COURSE 2	PYTHON ESSENTIALS FOR DATA SCIENCE	
MODULE 1	PYTHON INSTALLATION ANDSETUP	INSTALLING ANACONDA ON LOCAL MACHINE
		ADDITIONAL PACKAGE INSTALLATION
		INTRODUCTION TO JUPYTER NOTEBOOK
		JUPYTER NOTEBOOK KEYBOARD SHORTCUTS
MODULE 2	GOOGLE COLAB	INTRODUCTION TO GOOGLE COLAB
		SELECTION RUNTINE ENVIRONMENT GPU/TPU
		UPLOADING DATA/FILES IN COLAB
		LOADING GOOGLE DRIVE AS FOLDER
		SHARING COLAB NOTEBOOK
MODULE 3	PYTHON INTRODUCTION	NATIVE DATA TYPES
		KEY PYTHON FUNCTIONS
		SLICING OPERATIONS
		IMPORTING PACKAGES - DATETIME
		PACKAGE, SUB-PACKAGE, METHODS & ATTRIBUTES
MODULE 4	PYTHON DATA STRUCTURES	DATA STRUCTURES INTRODUCTION
		LISTS, LIST OPERATIONS
		TUPLE
		SETS
		DICTIONARIES
		LOOPING THROUGH ITERABLE DATA SET
		IF - CONDITIONAL STATEMENT

MODULE 5	PYTHON CONTROL STATEMENTS	FOR – LOOP
		USER DEFINED FUNCTIONS

COURSE 3	**R LANGUAGE ESSENTIALS**	
MODULE 1	R INTRODUCTION	R INSTALLATION AND SETUP
		R STUDIO – R DEVELOPMENT ENVIRONMENT
		R LANGUAGE BASICS
MODULE 2	R DATA SCIENCE	R DATA STRUCTURES
		R CONTROL STATEMENTS
		R DATA SCIENCE PACKAGES EXPLORATION
		PROJECT IN R

COURSE 4	**MATH FOR MACHINE LEARNING**	
MODULE 1	EIGENVALUES AND EIGENVECTORS	CALCULATING EIGENVALUES AND EIGENVECTORS
		EIGEN DECOMPOSITION OF A MATRIX
		EIGENVECTORS. WHAT ARE THEY?
MODULE 2	LINEAR TRANSFORMATIONSAND MATRICES	DETERMINANTS
		INVERSE
		RANK
		COLUMN AND NULL SPACE
		LINEAR TRANSFORMATIONS
		MATRICES: THE BASICS MATRIX OPERATIONS
		SYSTEM OF LINEAR EQUATIONS
MODULE 3	MULTIVARIABLE CALCULUS	CRITICAL POINTS
		MAXIMA AND MINIMA
		DIFFERENTIATION
		FUNCTIONS AND DERIVATIVES
		FUNCTIONS: PRIMER
		MULTIVARIABLE FUNCTIONS
		TAYLOR SERIES
		THE HESSIAN
		THE JACOBIAN
		VECTOR-VALUED FUNCTIONS
		DOT PRODUCT - EXAMPLE
		APPLICATION

MODULE 4	VECTORS AND VECTOR SPACES VECTOR-VALUED FUNCTIONS	INTRODUCTION TO LINEAR ALGEBRA
		VECTOR OPERATIONS - THE DOT PRODUCT
		VECTOR SPACES VECTORS: THE BASICS

COURSE 5	STATISTICS FOR DATA SCIENCE	
MODULE 1	INTRODUCTION TOSTATISTICS	DESCRIPTIVE AND INFERENTIAL STATISTICS. DEFINITIONS
		TERMS
		TYPES OF DATA
MODULE 2	HARNESSING DATA	TYPES OF SAMPLING DATA. SIMPLE RANDOMSAMPLING
		STRATIFIED
MODULE 3	EXPLORATORY ANALYSIS	MEAN
		MEDIAN AND MODE
		DATA VARIABILITY
		STANDARD DEVIATION
		Z-SCORE
		OUTLIERS
MODULE 4	DISTRIBUTIONS	NORMAL DISTRIBUTION
		CENTRAL LIMIT THEOREM
		HISTOGRAM
		NORMALIZATION
		NORMALITY TESTS
		SKEWNESS
		KURTOSIS.
MODULE 5	HYPOTHESIS TESTING	UNDERSTANDING HYPOTHESIS TESTING
		NULL AND ALTERNATE HYPOTHESES
		MAKING A DECISION
	HYPOTHESIS TESTING - CRITICAL VALUE METHOD	CRITICAL VALUE METHOD
		CRITICAL VALUE METHOD – EXAMPLES
	HYPOTHESIS TESTING – P-VALUE METHOD	P-VALUE METHOD
		P-VALUE METHOD – EXAMPLES
		TYPES OF ERRORS
	T-TESTS	T DISTRIBUTION
		ONE SAMPLE T-TEST
		INDEPENDENT AND RELATIONAL TWO-SAMPLE TEST
		T-TEST HYPOTHESIS TESTING IN PYTHON.
	ONE WAY ANOVA TEST / F-TEST	ANALYSIS OF VARIANCE (ANOVA) THEORY
		HYPOTHESIS TESTING WITH MORE THAN TWOVARIABLES WITH ANOVA
		INDUSTRY EXAMPLE
		F-TEST HYPOTHESIS TESTING IN PYTHON.
	NON-PARAMETRIC	CHI-SQUARE TEST THEORY

	HYPOTHESIS TESTING	APPLICATION OF CHI-SQUARE IN PYTHON
MODULE 6	CORRELATION & REGRESSION	DIRECT AND INDIRECT CORRELATION
		CORRELATION WITH STRONG AND WEAKCOLLERATION
		CALCULATING CORRELATION WITH PYTHON
		REGRESSION THEORY
		SIMPLE LINEAR REGRESSION WITH PYTHON

COURSE 6	DATA PREPARATION WITH NUMPY & PANDAS	
MODULE 1	NUMPY NUMERICAL PYTHONPACKAGE	INTRODUCTION
		NUMPY BASICS
		CREATING NUMPY ARRAYS
		STRUCTURE AND CONTENT OF ARRAYS
		SUBSET
		SLICE
		INDEX AND ITERATE THROUGH ARRAYS
		MULTIDIMENSIONAL ARRAYS
		PYTHON LISTS VS NUMPY ARRAYS
MODULE 2	OPERATIONS ON NUMPYARRAYS	BASIC OPERATIONS
		OPERATIONS ON ARRAYS
		BASIC LINEAR ALGEBRA OPERATIONS
MODULE 3	PANDAS PANEL DATA PACKAGE	PANDAS BASICS
		INDEXING AND SELECTING DATA
		MERGE AND APPEND
		GROUPING AND SUMMARIZING DATAFRAME
		LAMBDA FUNCTION & PIVOT TABLES
MODULE 4	DATA CLEANING DATA MUNGING WITHPANDAS	PANDAS BASICS
		INDEXING AND SELECTING DATA
		MERGE AND APPEND
		GROUPING AND SUMMARIZING DATAFRAME
		LAMBDA FUNCTION & PIVOT TABLES

COURSE 7	VISUALIZATION WITH PYTHON	
MODULE 1	BASICS OF VISUALIZATION	COMPONENTS OF A PLOT
		DATA VISUALIZATION TOOLKIT
		FUNCTIONALITIES OF PLOTS
		SUB-PLOTS
		INTRODUCTION
		PLOTTING AGGREGATE VALUES ACROSS CATEGORIES

MODULE 2	PLOTTING CATEGORICAL AND TIME SERIES DATA	PLOTTING DISTRIBUTIONS ACROSS CATEGORIES
		BIVARIATE DISTRIBUTIONS - PLOTTING PAIRWISERELATIONSHIPS
		VECTOR SPACES
		VECTORS: THE BASICS
MODULE 3	PLOTTING DATA DISTRIBUTIONS	INTRODUCTION
		UNIVARIATE DISTRIBUTIONS
		UNIVARIATE DISTRIBUTIONS - RUG PLOTS

COURSE 8	MACHINE LEARNING ASSOCIATE	
MODULE 1	MACHINE LEARNING INTRODUCTION	WHAT IS ML? ML VS AI
		ML WORKFLOW
		STATISTICAL MODELING OF ML
		APPLICATION OF ML
MODULE 2	MACHINE LEARNING ALGORITHMS	POPULAR ML ALGORITHMS
		CLUSTERING
		CLASSIFICATION AND REGRESSION
		SUPERVISED VS UNSUPERVISED
		CHOICE OF ML ALGORITHMS
MODULE 3	SIMPLE LINEAR REGRESSION	REGRESSION LINE
		BEST FIT LINE
	LINEAR REGRESSION IN PYTHON	ASSUMPTIONS OF SIMPLE LINEAR REGRESSION
		READING AND UNDERSTANDING THE DATA
		HYPOTHESIS TESTING IN LINEAR REGRESSION
		BUILDING A LINEAR MODEL
		RESIDUAL ANALYSIS AND PREDICTIONS
		LINEAR REGRESSION USING SKLEARN
	MULTIPLE LINEAR REGRESSION	SIMPLE LINEAR REG VS MULTIPLE LINEAR REG
		MULTICOLLINEARITY
		DEALING WITH CATEGORICAL VARIABLES
		MODEL ASSESSMENT AND COMPARISON
		FEATURE SELECTION
	LOGISTIC	INTRODUCTION: UNIVARIATE LOGISTICREGRESSION
		BINARY CLASSIFICATION

	REGRESSIONBINARY CLASSIFIER	SIGMOID CURVE
		FINDING THE BEST FIT SIGMOID CURVE SUMMARY
MODULE 4		MULTIVARIATE LOGISTIC REGRESSION
		DATA CLEANING AND PREPARATION
		BUILDING YOUR FIRST MODEL
	LOGISTIC REGRESSIONMODEL BUILDING	FEATURE ELIMINATION USING RFE
		CONFUSION MATRIX AND ACCURACY
		MANUAL FEATURE ELIMINATION

COURSE 8	MACHINE LEARNING ASSOCIATE	
MODULE 4	LOGISTIC REGRESSIONMODEL EVALUATION	METRICS BEYOND ACCURACY: SENSITIVITY &SPECIFICITY
		FINDING THE OPTIMAL THRESHOLD USING ROCCURVE
		METRICS BEYOND ACCURACY: PRECISION & RECALL
MODULE 5	SUPERVISED LEARNING:K NEAREST NEIGHBOR KNN CLASSIFIER	INTRODUCTION TO KNN
		HOW IT WORKS: THEORY
		PROS AND CONS OF KNN
		APPLICATIONS OF KNN
		MODEL BUILDING KNN IN PYTHON SKLEARN
		EVALUATION: KNN MODEL.
MODULE 6	UNSUPERVISED LEARNING:CLUSTERING	INTRODUCTION
		UNDERSTANDING CLUSTERING
		PRACTICAL EXAMPLE OF CLUSTERING - CUSTOMERSEGMENTATION
	K MEANS CLUSTERING	INTRODUCTION
		STEPS OF THE ALGORITHM
		K MEANS ALGORITHM
		K MEANS AS COORDINATE DESCENT
		VISUALISING THE K MEANS ALGORITHM
		PRACTICAL CONSIDERATION IN K MEANSALGORITHM
		CLUSTER TENDENCY
	K MEANS IN PYTHONCASE: IRIS DATASET	INTRODUCTION
		IRIS DATA PREPARATION
		MAKING THE CLUSTERS

	CLUSTERING	
MODULE 7	HIERARCHICAL CLUSTERING	INTRODUCTION
		HIERARCHICAL CLUSTERING ALGORITHM
		INTERPRETING THE DENDROGRAM
MODULE 8	UNSUPERVISED LEARNING:PRINCIPLE COMPONENT ANALYSIS (PCA)	THE WHY'S AND WHAT'S OF PCA
		BUILDING BLOCKS OF PCA
		ILLUSTRATION - FINDING PRINCIPAL COMPONENTS
		COMPREHENSION - CALCULATING THE PRINCIPALCOMPONENTS
		SINGULAR VALUE DECOMPOSITION (SVD)

COURSE 9	ADVANCED MACHINE LEARNING	
Module 1	CLASSIFICATION AND REGRESSION TREE (CART):DECISION TREE	INTRODUCTION TO DECISION TREES
		INTERPRETING A DECISION TREE
		COMPREHENSION - DECISION TREE CLASSIFICATION INPYTHON
		REGRESSION WITH DECISION TREES
Module 2	THEORY OF DECISION TREE	INTRODUCTION
		CONCEPT OF HOMOGENEITY
		GINI INDEX
		ENTROPY AND INFORMATION GAIN
		COMPREHENSION - INFORMATION GAIN
		SPLITTING BY R-SQUARED
Module 3	DECISION TREE HYPER-PARAMETER TUNING	BUILDING DECISION TREES IN PYTHON
		CHOOSING TREE HYPERPARAMETERS IN PYTHON
		COMPREHENSION - HYPERPARAMETERS
		TREE TRUNCATION
		ADVANTAGES AND DISADVANTAGES TREE TRUNCATION
Module 4	RANDOM FOREST ENSEMBLE BAGGING TECHNIQUE	INTRODUCTION
		ENSEMBLES
		COMPREHENSION - ENSEMBLES
		CREATING A RANDOM FOREST
		COMPREHENSION - OOB (OUT-OF-BAG) ERROR
		RANDOM FORESTS LAB
	NAÏVE BAYES: BAYES THEOREM AND	INTRODUCTION: NAIVE BAYES
		CONDITIONAL PROBABILITY AND ITS INTUITION
		BAYES' THEOREM
		NAIVE BAYES WITH ONE FEATURE

	ALGORITHMBUILDING BLOCKS	CONDITIONAL INDEPENDENCE IN NAIVE BAYES
		DECIPHERING NAIVE BAYES
Module 5	NAÏVE BAYES: TEXT CLASSIFICATION HAM VS SPAM CASESTUDY	INTRODUCTION NAIVE BAYES FOR TEXTCLASSIFICATION
		DOCUMENT CLASSIFIER PRE-PROCESSING STEPS
		DOCUMENT CLASSIFIER WORKED OUT EXAMPLE
		LAPLACE SMOOTHING
		BUILDING SPAM HAM CLASSIFIER
		COMPREHENSION NAIVE BAYES FOR TEXT CLASSIFICATION

COURSE 9	ADVANCED MACHINE LEARNING	
Module 6	BOOSTING: INTRODUCTION, ADABOOST, GRADIENT BOOSTING, XGBOOST	INTRODUCTION TO BOOSTING
		WEAK LEARNERS
		ADABOOST ALGORITHM
		ADABOOST DISTRIBUTION AND PARAMETER CALCULATION
		ADABOOST LAB
		UNDERSTANDING GRADIENT BOOSTING
		GRADIENT IN GRADIENT BOOSTING
		GRADIENT BOOSTING ALGORITHM
		XGBOOST
		KAGGLE PRACTICE EXERCISE
Module 7	SUPPORT VECTOR MACHINE:THEORY	INTRODUCTION TO SVM
		CONCEPT OF A HYPERPLANE IN 2D
		CONCEPT OF A HYPERPLANE IN 3D
		MAXIMAL MARGIN CLASSIFIER
		THE SOFT MARGIN CLASSIFIER
		THE SLACK VARIABLE
		NOTION OF SLACK VARIABLES
		COST OF MISCLASSIFICATION
	SVM : IMPLEMENTING SVM INSKLEARN, CASESTUDY	MAPPING NONLINEAR DATA TO LINEAR DATA
		FEATURE TRANSFORMATION
		THE KERNEL TRICK
		MODELING SVM PYTHON SKLEARN
		MODEL EVALUATION
		INTRODUCTION TO ANN
		SIMPLE ANN NETWORK
		HOW IT WORKS: BACKPROP ALGORITHM

Module 8	ARTIFICIAL NEURAL NETWORK(ANN)	IMPLEMENTING ANN WITH PYTHON SKLEARN
		ANN MODELING AND EVALUATION
		COMPREHENSION
Module 9	ADVANCED ML CONCEPTS	ADV EVALUATION METRICS: ROC_AUC, R2 THEORY,PRECISION, RECALL, F1 SCORE, RMSE
		K-FOLD CROSSVALIDATION
		GRID AND RANDOMIZED SEARCH CV IN SKLEARN
		IMBALANCED DATA SET : SMOTE TECHNIQUE
		FEATURE SELECTION TECHNIQUES
		CHOOSING RIGHT ALGORITHMS

COURSE 10	SQL FOR DATA SCIENCE	
MODULE 1	INSTALL SQL PACKAGES AND CONNECTING TO DB	SQLALCHEMY
		PYMYSQL
MODULE 2	RDBMS (RELATIONAL DATABASE MANAGEMENT) BASICS	BASICS OF SQL DB
		PRIMARY KEY
		FOREIGN KEY
MODULE 3	SELECT SQL COMMAND, WHERE CONDITION	RETRIEVING DATA WITH SELECT SQL COMMAND
		WHERE CONDITION TO PANDAS DATA FRAME.
MODULE 4	ADVANCED SQL	ORDER BY CLAUSE
		AGGREGATE FUNCTIONS
		GROUP BY CLAUSE
		HAVING CLAUSE
		NESTED QUERIES
		INNER JOIN, OUTER JOINS, MULTI JOIN

COURSE 11	DEEP LEARNING – CNN FOUNDATION	
MODULE 1	INTRODUCTION TO DEEP LEARNING	WHAT IS DEEP LEARNING?
		VARIOUS DEEP LEARNING MODELS IN PRACTICEAND APPLICATIONS
MODULE 2	INTRODUCTION TO IMAGEBASICS	IMAGE RESOLUTION
		PIXELS
		IMAGE MANIPULATIONS WITH FILTERS
MODULE 3	CONVOLUTIONAL NEURALNETWORK CNN INTRO	CNN ESSENTIALS
		CNN ARCHITECTURE
		WORK FLOW OF IMAGE CLASSIFICATION WITHCNN

MODULE 4	CASE STUDY: KERAS– TENSORFLOW IMAGE CLASSIFICATION	CNN HANDS ON APPLICATION FOR CLASSIFICATION IMAGES OF CATS AND DOGS

COURSE 12	**TABLEAU ASSOCIATE**	
MODULE 1	TABLEAU INTRODUCTION	TABLEAU INTERFACE
		DIMENSIONS AND MEASURES
		FILTER SHELF
MODULE 2	CONNECTING TO DATA SOURCE	CONNECTING TO SOURCES
		EXCEL
		DATABASE
		PDF
MODULE 3	VISUAL ANALYTICS	CHARTS AND PLOTS WITH SUPERSTORE DATA
MODULE 4	FORECASTING	FORECASTING TIME SERIES DATA
		FORECASTING SALES IN TIME PERIODS

COURSE 13	**MACHINE LEARNING MODEL DEPLOYMENT - API**	
MODULE 1	BASICS OF APPLICATION PROGRAM INTERFACE (API)	API BASICS
		LOOSELY COUPLED ARCHITECTURE
MODULE 2	INSTALLING FLASK ANDFLASK CORS	INSTALLATION AND CONFIGURING FLASK
		CROSS DOMAIN AUTHENTICATION WITHFLASK_CORS
		EXAMPLE TO USE FLASK AS API SERVER
MODULE 3	END OF END ML PROJECTWITH API DEPLOYMENT	COMPLETE PROJECT FLOW WITH API DEPLOYMENT AND ASSESSING THROUGHWEBSITE

COURSE 14	**BIG DATA ESSENTIALS**	
MODULE 1	BIG DATA INTRODUCTION	WHAT IS BIG DATA?
		VARIOUS BIG DATA FRAMEWORKS
		HADOOP INTRODUCTION

MODULE 2	HADOOP AND SPARK	SPARK BIG DATA FOR MACHINE LEARNING
		MANAGING BIG DATA IN DATA SCIENCEPROJECTS

COURSE 15	DATA SCIENCE PROJECT EXECUTION	
MODULE 1	DATA SCIENCE: PROJECTSTRUCTURE	CRISP DM FRAMEWORK
		6-PHASE PROJECT EXECUTION
MODULE 2	BUSINESS ASPECTS	ML USE CASE DEVELOPMENT
		PROJECT MANAGEMENT METHODOLOGY
		CHALLENGES AND PITFALLS