CODING CREATROS

Data Science Course Content

Contents

Data Science Course Content

1. Introduction to data science

1. What is Data Science
2. How it is different from Big Data and Data Analytics
3. Data Driven decision making
4. Purpose and Business problems
5. How Data Scientist work
6. Skills of a data scientist
7. Different sectors using Data science
8. Real World Applications
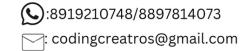9. Future of AI and how the world is changing

2. Statistics

- Introduction to Statistics
  - Statistical and Non-Statistical Analysis
  - Major categories of statistics – Frequency and Bayesian
  - Difference between Statistics and Probabilities
  - Statistical terms
  - Difference between Descriptive Statistics and Inferential Statistics
  - Understanding of Population and Samples
- Descriptive Statistics
- Inferential Statistics
- Central Limit Theorem
- **Types of variables**
  - Nominal/Categorical
  - Ordinal
  - Interval/Ratio
  - Continuous, Time Series
- **Central Tendency**
  - Mean
  - Median
  - Mode
- **Measure of Statistical dispersions**
  - Variance and Bessel correction
  - Standard Deviation
  - Standard Error
  - Margin of Error
  - IQR
  - Range
  - Mean absolute difference
  - median absolute deviation
  - Coefficient of variance
  - Skewness
  - Law of Large Numbers
  - Confidence Level & Interval
  - P value and its interpretation
  - Correlation and auto correlation & correlation matrix
  - Correlation ratio

Data Science Course Content

- Sampling Techniques
- Sampling errors
- Sample size estimation
- Point estimation & margin of error
- Multi Collinearity
- Co-variance and correlation
- P- value and critical value approach
- T-Distribution and T-Statistics
- **Hypothesis testing's**

  - What is Hypothesis Testing Different types of Errors
  - (Type I and Type II Errors) Z-test T-test Chi-square
  - test ANOVA (one way and two way) F-test & f score
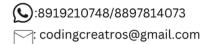 - P-Value & Significance Level
  -
  -
  -
  -

3. Probability
- Probability
- Venn diagram
- counting (permutation & combination)
- Expectation
- Rules of Probabilities
- Bayesian Network
- Random Variables and Expected Values
- Bayes theorem
- Maximum likelihood estimation
- Probability Distributions

  - Continuous Distributions- (Normal, uniform, T, F, chi square)
  - "Discrete Distributions- (Bernoulli, binomial, Poisson)
  - Empirical Rules with Z- Score

4. Python
- Why python for data analysis
- how to install Anaconda
- Running few simple programs using python
- "Python objects
  - Lists
  - Strings
  - Tuples
  - Dictionaries"
  - Arrays, Data frames in python
- "Python Libraries
  - Numpy
  - Scipy
  - Matplotlib
  - Pandas
  - Scikit Learn
  - Seaborn
  - regular expressions

- Introduction to Series and Data frames
- Math functions
- User defined Functions
- Parameters and arguments of functions
- Recursive function and its examples
- "Conditionals in python

  - o If loop
  - o elif
  - o if elif else"
  - o "Loops in python
  - o for loop
  - o while loop"
- Introduction   to   pandas
- Broadcasting   in   Python
- Array shape manipulations
- Data structures in pandas

  - o Series
  - o Data frame
  - o Panel"
- "Various Data Frame Operations
  - o Selection
  - o Deletion etc.
  - o "Grouping, Merging, and Reshaping of Data
- Creating matrixes using numpy
- Statistical operators using Numpy

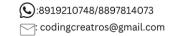## 5. Data and Data Science Thinking

- Basics of data categorization and different formats of data
  - o Structured Data
  - o Unstructured Data
  - o Time Series
- Why and how to raise the right question
- Correlation is not the causation and its importance
- Limitations as a data scientist
- Transformation of intuition-based decision making to data driven
- Story Telling

6. Data Analytics Overview

- Data Analytics Process
- Exploratory Data Analysis(EDA)
- How to start with Data Analytics Project
- Intro to Web Scrapping and Beautiful Soup

7. Machine Learning, Data Science and Artificial Intelligence

- Supervised Learning
- Unsupervised Learning
- Difference between Classification and Regression
- **Data pre-processing**

  - o What is data set.
  - o What is training set
  - o What is test set and need for test set

- o Expectation-Maximization technique for missing value
- o using Gradient
- o Feature scaling
- o binning
- o one hot encoding
- o Feature engineering
- o Outliers treatment
- o Bias and Variance trade off
- o Over fitting and Under fitting

- **Exploratory Data analysis(EDA)**
  - o Univariate analysis
  - o Bivariate Analysis
  - o Feature Engineering
  - o Variable transformation
  - o Variable /Feature Creation
  - o Project

- **Supervised Regression Algorithms**
  - o Simple Linear Regression
  - o Multiple Linear Regression
  - o Ordinary Least Square(OLS)
  - o Decision tree Regression
  - o Random Forest Regression
  - o GLM (Poisson regression, spline)
  - o Support Vector Machines Regression
  - o Error and Accuracy
  - o Gradient Descent
  - o Regularization Techniques
  - o Maximum Likelihood estimation(MLE)
  - o Probabilistic diagnosis of outliers
  - o L2 and L1 Norms
  - o Ridge Regression
  - o Lasso Regression and ElasticNet
  - o Project

- **Supervised Classification Algorithms**
  - o Logistic regression classification
  - o Multiclass Classification using Logistic Regression
  - o Decision tree Classification
  - o Random Forest classification
  - o Support Vector Machines classification
  - o What is Naïve Bayes theorem and the limitation
  - o Naïve Bayes Classification
  - o Ada boost/ Adaptive - Boosting Algorithm
  - o GBM
  - o Probability in Classification
  - o Creating the log loss formula with entropy
  - o Softmax Function
  - o MLE in classification
  - o Understanding the Neural Networks
  - o SVM
  - o Gradient Boosting
  - o XG Boost (Extreme Gradient Boosting)
  - o Project

Data Science Course Content

- **Unsupervised Algorithms**
  - o K-means Clustering
  - o Hierarchical clustering
  - o Association Rule Mining
  - o KNN Classifier
  - o PCA
  - o Project

  - **Model Evaluation Metrics**
  - o ROC Curves
  - o Confusion matrix
  - o Accuracy
  - o Recall & Precision
  - o Specificity & Sensitivity
  - o Receiver Operating Characteristic (ROC) curve
  - o Area Under Curve (AUC)
  - o F1-Score
  - o AIC & BIC Scores
  - o R squared & Adjusted R squared
  - o RMSE, MSE

- **Model selection Techniques**
  - o Cross validation
  - o Bootstrap
  - o Model selection using Statistical tests
  - o Grid search
  - o Evaluation Matrix
- **Natural Language Processing (NLP)**
  - o What is NLP
  - o Cleaning Text
  - o Tokenization
  - o Term Frequency (TF)
  - o Term Frequency – Inverse Document Frequency (TF-IDF)
  - o Document Term Matrix

- **AI and Deep Learning**
  - o Introduction to Deep Learning and Neural Networks
  - o Introduction to Linear Algebra
  - o Artificial Neural Networks
  - o Activation Functions
  - o Back Propogation
  - o Chain Rule of Differentiation
  - o Vanishing Gradient Descent
  - o Exploding Gradient Descent
  - o Drop Out Layers in Multi Neural Network
  - o Deep Learning-Activation Functions-Elu, PRelu,Softmax,Swish And Softplus
  - o Weight Initialization Techniques
  - o Gradient Descent vs Stochastic Gradient Descent
  - o AdaGrad Optimizers
  - o Hyper Parameter Tuning
  - o CNN
  - o CNN vs ANN
  - o LSTM
  - o Bi-LSTM

- **Generative AI**

Data Science Course Content

o   Introduction to Generative AI
o   Introduction to Langchain
o   Memory in Langchain
o   Introduction to Vector Database for AI &Large Language Models (LLM)