Kelly Technologies

DATA SCIENCE (AI, ML, DL)

Foundation
Data Analysis, Visulaization & Programming – 3 Modules

Module – 1: Business Statistics & Visualization

Module – 2: Python Programming (Numpy/Pandas/Matplotlib/Sci-Kit Learn

Module – 3: Hands-On-Project (Statistics & Data Visualization)

Specialization
Data Science – Machine Learning & Artificial Intelligence
4 Modules

Module – 4 : Machine Learning

Module – 5 : Artificial Intelligence / Deep Learning

Module – 6   : Deploying Models on Production

Module – 7 : Hands-On-Project (ML & AI)

1. **Introduction to Data Science**
    a. What is Data Science?
    b. Difference between Data Science, Artificial Intelligence, Machine Learning, Deep Learning, Generative AI
    c. Who are Data Scientists? Roles, Job Descriptions, Pay Scales ?
    d. Real-Time Data Science Project Life Cycle
    e. Data Science Project Architecture

Module – 1	Business Statistics & Data Visualization

2. Business Statistics

a. Qualitative & Quantitative
b. Scales of Measurement – Nominal / Ordinal / Interval / Ratio
c. Types of Statistics
- Descriptive Statistics
- Inferential Statistics

d. *Descriptive Statistics*

- Measures of Central Tendency or Location
  - Mean
  - Median
  - Mode
  - Percentiles
  - Quartiles (Q1 – 25th Percentile, Q2 – 50th Percentile/Median, Q3 – 75th Percentile)

e. Measures of Dispersion or Spread or Variability
- Variance
- Standard Deviation
- Interquartile Range
- Range
- Coefficient of Variation

f. Measures of Distribution – Shape
- Distribution Shape
  - Histogram (Summarizing Quantitative Data)
  - Skewness – Left & Right Skewed
- Normal Distribution / Normal Curve / Normal Probability Distribution
  - Bell Shaped Curved
  - Understanding Properties of Normal Distribution
  - Area Under the Curve for any Normal Distribution
  - Normal Probability Density Function
  - Standard Normal Distribution
  - Z – scores – Standard Normalization (computation)

g. *Inferential Statistics*

- Point Estimation
  - Population & Sample
    - ✓ Population Parameter

- ✓ Sample Statistics
- ✓ Difference between Population Parameter & Sample Statistic?
- ✓ Why Sample when Population data available?
- ✓ Sample Size to select
- ✓ Determining sample size for a hypothesis test about a population mean
- ✓ Sampling Methods or Distributions
  - o Simple Random Sample
  - o Systematic Sampling
  - o Stratified Sampling
- ▪ Hypothesis Testing and Decision Making
  - • What is hypothetical question (more than one answer)
  - • Developing Null and Alternate Hypothesis
  - • Alternate Hypothesis as a Research Hypothesis and we need evidence to believe
  - • Null Hypothesis is an assumption to be challenged
  - • Confidence Level
  - • Type I & Type II Errors
  - • T-Test (comparison of means)
  - • Comparison between two groups
    - ✓ One Sample t-test
      - o One Sided/Tailed
      - o Two Sided/Tailed
    - ✓ Two Sample t-test
      - o One Sided/Tailed
      - o Two Sided/Tailed
    - ✓ Paired t-test
  - • Steps for Evaluating Hypothesis – One Sample & Two Samples
    - ✓ Compute confidence level with t-test
    - ✓ Finding out t-critical values from t-distribution table
    - ✓ Degrees of Freedom
    - ✓ Calculating p-value (probability of making Type I Error)
- h. Measures of Association between Two Continuous Variables
  - ▪ Covariance
  - ▪ Correlation Coefficient (Measure of Relationship)
  - ▪ Interpretation of Covariance & Correlation Coefficient
  - ▪ Cause & Effect relationship
  - ▪ Difference between spurious correlation & correlation

i. Probability
- Sample Space
- Event
- How do we calculate Probability?
  - Classical approach
  - Relative frequency approach
  - Conditional Probability
  - Bayesian Conditional Probability

j. Discrete Distributions
- Ratio & Proportions
- Odds & Odds Ratio
- Conditional Odds Ratio
- Discrete Distributions on Binary Variable
  - Binomial Distribution
  - Poisson Distribution
  - Uniform Distribution
- Test of Independence
  - Chi-Square test for association (Measure of Relationship between two categorical/binary variables)
  - Chi-Square distribution
  - Degrees of Freedom in Chi-Square statistics
  - Chi-Square statistic formula
  - Steps for hypothesis testing using chi-square test

k. ANOVA – Analysis of Variance
- F-Test
  - Compute F-Statistics
  - F-Statistics Formula (between groups and within the group)
- Compute F-Critical value
- Degrees of Freedom
- Confidence
- Steps for hypothesis testing using F-statistics and F-critical values for more than two groups

3. Data Pre Processing
a. Cleaning data with python
b. Data Type Conversions
c. Encoding categorical data
d. Binning and Normalization

e. Feature Scaling & Standardizing Data
f. Handling missing values – Imputation

4. Exploratory Data Analysis (EDA) & Visualization
a. Summary Statistics
b. Charts & Graphs
- One Dimensional Charts
    - Histogram / BarChart
- Two Dimensional Charts
    - Bar Charts (Stack & Dodge)
    - Box Plots
    - Scatter Plots
- Multi-Dimensional Plots
- Fancy charts – Bubble Charts, Word Clouds
- Outlier Detection & Management
- Variable Selection / Variable Transformation

Module – 2   Python Programming (Numpy/Pandas/Matplotlib/ Sci-Kit Learn)

---

1. Introduction to Python
a. Understanding the Reason for Python's popularity
b. Different IDE, Anaconda and Jupyter Notebook
c. **Basic, Core & Advanced Python:**
- Variables & Data Types in Python
- Functions
- Local & Global Statements
- Data Structures: List, Tuple, Set, Dictionaries
- Operators, Loops, Functions, Dictionaries
- Numeric & String related functions
- Object Oriented Programming (OOP)
d. **Scientific Python:**
- Numpy
- Pandas
- Matplotlib
- Seaborn
- Sci-kit Learn

Module – 4   Machine Learning

---

5. Machine Learning Introduction

   a. Introduction to Machine Learning

   b. Types of Machine Learning Algorithms:
- Supervised Learning
- Unsupervised Learning

   c. Selection of algorithm based on problem statement

   d. Difference between Regression & Classification

   e. Splitting data into
- Train, Test & Validation dataset

   f. Measures of Performance
- R-Square
- RMSE for Regression
- Confusion Matrix
- Accuracy, Precision & Recall
- F-1 Score
- Sensitivity & Specificity
- Roc & Auc

Supervised Learning

6. Linear Regression (Simple & Multiple)

   a. Introduction to Linear Regression
- Simple Linear Regression
- Multiple Linear Regression

   b. Model development, improvement and interpretation

   c. Sum of Least Squares

   d. Cost Function

   e. Feature Selection

   f. Model validation & diagnostics – Tests to validate assumptions

   g. Disadvantages of linear models

   h. Regularization Techniques
- Ridge Regression (L2 Regularization)
- Lasso Regression (L1 Regularization)

   i. Non-Linear Regression

7. Logistic Regression

a. Variable Selection Methods
✓ Forward, Backward & Stepwise
b. Need for logistic regression
c. Logit link function
d. Maximum Likelihood Estimation
e. Model Development And Validation
f. Measures of Performance – Confusion Matrix / Roc Curve / RMSE / MAPE / Misclassification Rate / AUC / Precision & Recall / Decile / Lift & Gain Charts
g. Advantages and disadvantages of logistic regression models

8. Decision Trees
a. Decision Trees Types
- Regression Tree
- Classification Tree
b. C5.0
c. Classification and Regression Trees (CART)
d. Rule based learning
e. Construction of Rules
f. Choosing variables for decision nodes / Feature Importance
g. Over-fitting and Pruning
h. Setting constraints – Number of Samples, Features & Trees / Size / Depth of Tree etc.
i. Entropy / Gini / Chi-Square / Information Gain
j. Advantages of tree based models

9. Bagging & Random Forest
a. Bagging Technique and need of Bagging
b. Re-sampling Methods with and without Replacement
c. Hyper Parameter Tuning

10. Boosting
a. Introduction to Boosting Techniques
b. Types of Boosting
- Gradient Boosting Machine (GBM)
- Extreme Gradient Boosting Machine (XGBM)
- AdaBoost
c. Application of Boosting in Classification and Regression Models

11. Cross Validation

a. Leave One Out Cross Validation
b. K-Fold Cross Validation
c. Bias & Variance

12. Forecasting – Time Series Analysis

a. Components of Time Series
   ▪ Trend / Seasonality / Randomness
b. Additive & Multiplicative
c. Simple moving averages
d. Exponential smoothing
   ▪ Dousie & Triple Exponential Smoothing
e. Time series decomposition
f. Holt Winters
g. ARIMA & ARIMAX

13. Support Vector Machines (SVM)

a. Maximum Margin Classifier
b. Support Vector Classifier
c. Kernels – Linear and Non Linear

UnSupervised Learning

14. Clustering (Segmentation)

a. Hierarchical Clustering
b. K-Means Clustering
c. Cluster Profiling
d. KNN

15. Dimensionality Reduction Techniques

a. Principal Component Analysis (PCA)

16. Association Rules (APRIORI Algorithm)

17. Recommender Systems

a. Collaborative Filtering
b. Non-Negative Matrix Factorization
c. Anomaly Detection

Module – 5	Artificial Intelligence – Deep Learning

18. Text Mining
a. Text Analytics
b. Unstructured Data
c. Natural Language Processing (NLP)
d. Cleaning Text Data: Tokenization, Pre-Processing
e. Sentiment Analysis
f. Text Classification & TFIDF
g. POS Tagging
h. Topic Modeling
i. Feature Extraction

19. Probabilistic Methods Introduction
a. Naïve Bayes
b. Joint & Condition Probabilities
c. Classification Using Naïve Bayes Approach

20. Artificial Neural Networks (ANN)
a. Perceptron
b. Single Layer Perceptron
c. Multi Layered Perceptron
d. Activation Functions & Types
   - ReLU / Sigmoid / Tanh
e. Forward & Backward Propagation
f. Training & Epochs
g. Dropout & Drop connect
h. Tools: TensorFlow & Keras

21. Convolution Neural Networks (CNN)
a. CNN for Computer Vision
b. CNN for Image Classification
c. CNN
   - Convolution Layers
   - Pooling Layers
   - Filters & Feature Maps

- ▪ Padding & Strides
- ▪ Training of CNN

22. Recurrent Neural Networks (RNN)
a. Sequential Networks
b. RNN Structure
c. Introduction to LSTM
d. Word Embedding
   - ▪ Glove
   - ▪ Word2Vec

Module – 6	Deploying Models on Production

Module – 7	Hands – On Projects (ML & AI-DL)

23. Data Science Project Life Cycle
a. Proof of Concept (POC)

Key Takeaways

a. Data Analytics Hands-On-Project delivered using Tools like: Excel ,Sql, PowerBi & Python
b. Data Science Hands-On-Project delivered on Machine Learning & Deep Learning using Tools like: Python, Keras & Tensorflow
c. Resume preparation
d. All topics explained with REAL WORLD projects only
e. Projects End to End Lifecycle explanation
f. Mock Interviews & Test for Data Science Interview preparation
g. Detailed assistance in Resume preparation. Special attention for experienced people on previous experience