**Data Science Workflow**

**Exploratory Data Analysis**

* 1. Continuous Data
     1. Measures of Central Tendency:
        1. Mean
        2. Median
        3. Mode
     2. Measure of Dispersion
        1. Variance
        2. Standard Deviation
        3. Coefficient of Variation
        4. Percentile
        5. Range
        6. Interquartile Range (IQR)
     3. Measures of Shape
        1. Skewness
        2. Kurtosis
  2. Discrete Data
     1. Count
  3. Graphical Representations
     1. Bar Chart
     2. Box Plot
     3. Histogram
     4. Pie Chart
     5. Quantile-Quantile Plot (Q-Q Plot)
     6. Scatter Plot
  4. Normality Test
     1. Pearson’s Chi-Squared Test
     2. Shapiro-Wilk Test
     3. D’Agostino’s Test
     4. Anderson-Darling Test
  5. Correlation Test
     1. Pearson’s Correlation Coefficient (Linear Relationship)
     2. Spearman’s Rank Correlation (Monotonic Relationship)
     3. Kendall’s Rank Correlation (Monotonic Relationship)
     4. Chi-Squared Test (Test for independence of two samples)
  6. Hypothesis Test (Significantly Different Test)
     1. Student’s t-test
     2. Analysis of Variance Test (ANOVA)
  7. Nonparametric Test (Significantly Different Test)
     1. Mann-Whitney U Test
     2. Wilcoxon Signed-Rank Test
     3. Kruskal-Wallis H Test
     4. Friedman Test

**Data** **Preparation**

* 1. Data Cleaning
     1. Inconsistent Data
     2. Missing Data
     3. Noisy Data (Outliers)
  2. Data Integration
     1. Conflicting Data
     2. Redundant Data
  3. Data Transformation
     1. Noisy Data
     2. Normalization
     3. Aggregation
     4. Generalization
  4. Data Reduction
     1. Dimensionality Reduction
     2. Data Cube Aggregation
     3. Numerosity Data
     4. Sampling
  5. Data Discretization
     1. Binning
  6. Train Test Split

**Model Building**

* 1. Regression:
     1. Linear Regression
     2. Logistic Regression
     3. Stepwise Regression
  2. Instance Based:
     1. K-Nearest Neighbors (kNN)
     2. Support Vector Machines (SVM)
  3. Regularization:
     1. Ridge Regression
     2. LASSO Regression
  4. Decision Tree
     1. Classification and Regression Tree (CART)
     2. Random Forest
     3. XG Boosting
  5. Bayesian
     1. Naïve Bayes
     2. Gaussian Naïve Bayes
     3. Multinomial Naïve Bayes
     4. Bayesian Belief Network
     5. Bayesian Network
  6. Clustering;
     1. K-Means
     2. K-Medians
     3. Hierarchical Clustering

1. **Model Validation / Data Wrangling**
   1. Regression:
      1. Goodness of Fit (r-squared & Adjusted r-squared)
      2. Analysis of Residuals
         1. Residuals vs Predictors (X variables)
         2. Fitted vs Residuals (homoscedastic, heteroscedastic, nonlinear)
         3. Response vs Fitted
      3. Accuracy Metrics
         1. Mean Absolute Error (MAE)
         2. Mean Squared Error (MSE)
         3. Root Mean Squared Error (RMSE)
         4. Mean Absolute Squared Error (RMSE)
      4. Normality
      5. Outlier Detection (Cook’s Distance Plot)
      6. Variance Influence Factors
      7. Heteroscedastic, Homoscedastic Test (Test for constant variance)
         1. Breusch-Pagan Test
         2. Bartlett’s Test
         3. White’s Test
   2. Classification
      1. Confusion Matrix
         1. Accuracy
         2. Misclassification Rate
         3. Sensitivity
         4. Specificity
         5. F1 Score
      2. Receiver Operator Characteristic (ROC)