Exploratory Data Analysis (EDA)

Get a feel for the data.

- What are your features?
- Are there missing data?
- Plot histograms, scatters, correlations, pair plots.
- Data cleaning: deal with missing data and/or outliers, format strings, etc.

Problem Formulation

What is the question you're asking or thesis you're looking to support?

Are you predicting a value? Regression
Are you predicting a category? Classification
Are you grouping unlabeled data? Clustering

A. Regression
B. Classification
C. Clustering, Dimensionality Reduction

A. Regression

- Linear
- Polynomial

Pre-Processing

Train-test split

Scaling
- StandardScaler
- MinMax
- PCA

Feature Engineering
- PolynomialFeatures

Encoding
- OrdinalEncoder
- OneHotEncoder
- LabelEncoder

Model Selection

Regularization
- Ridge
- Lasso
- ElasticNet

Parameter Search/ Validation
- GridSearchCV
- LinearRegressionCV
- RidgeCV
- ElasticNetCV

Model Fit
- fit, predict

Assessment

Metrics
- R^2
- Mean Squared Error
- Mean Absolute Error
- Coefficients (feature importance)

Plots
- Scatter with trend line

Check
- Histogram of residuals
- R^2 Training vs Testing

B. Classification

- LogisticRegression
- DecisionTreeClassifier
- RandomForestClassifier
- GradientBoostedClassifier
- KNeighborsClassifier

Pre-Processing

Train-test split

Scaling *(not for trees/forests)*
- StandardScaler
- MinMax
- PCA

Encoding
- OrdinalEncoder
- OneHotEncoder
- LabelEncoder

Model Selection

Regularization
- Logisitic - Ridge, Lasso, Elastic
- Trees – max_depth, min_split, others
- KNN – kneighbors

Parameter Search/ Validation
- GridSearchCV

Model Fit
- fit, predict

Assessment

Metrics
- Accuracy
- Recall
- Precision
- F1

Plots
- Confusion Matrix
- Plot tree
- Decision Boundary

Check
- Recall vs Precision
- Training vs Testing

C. Clustering, PCA

- KMeans
- PCA

Pre-Processing

Train-test split

Scaling *(not for trees/forests)*
- StandardScaler
- MinMax
- PCA

Encoding
- OrdinalEncoder
- OneHotEncoder
- LabelEncoder

Model Selection

Parameter Search/ Validation
- GridSearchCV
- Loop over params

Model Fit
- fit, transform

Assessment

Metrics
- Clustering – Silhouette score
- PCA – explained_variance_, singular_values_

Plots
- Clustering - Colored scatter (for 2d)
- PCA - Scree plot (line graph of explained variance)