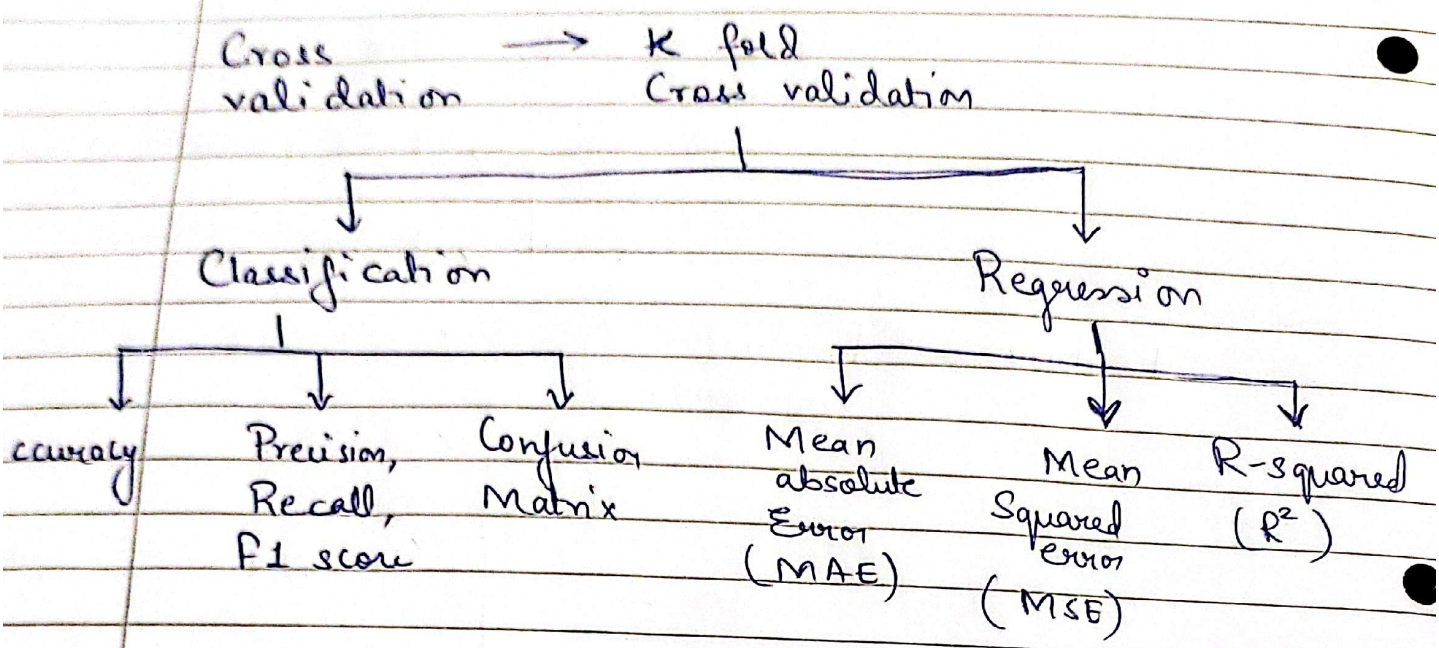
Data → Data Preparation
- → Data Cleaning
- → Data Transformation
- → Data Splitting

Feature → Feature Engineering
- → Feature selection
- → Feature creation
- → Feature Transformation
- → Dealing with missing/inconsistent data

Training & Testing set
- → 80 – 20 split b/w training data & testing data.
- → 70 – 30 split b/w training data & testing data.
- → 60 – 20 – 20 split b/w training, te & validation

* validation set
↓
helps tune hyper parameters without affecting dataset



Model & Algorithm

Algorithms → steps or procedures used to learn patterns from data.
They define how the model will adjust its internal parameters to improve its predictions

Model → o/p of the algorithm after training on data.

Example Project	Data Type	Feature Type	Label Type	Model Type	Algorithm Type	Training, Testing & Validation
Sentiment Analysis	Text	Word embeddings (eg. TF-IDF, BERT)	Categorical eg. (positive, negative)	Text Classification	**Naive Bayes:** Calculates the probability of each class based on feature occurence assuming feature independence. ~~Transformer~~ ↓ **Transformer (BERT):** Uses attention mechanisms to capture long range dependencies & word context	Train, test split with k fold Cross validation Transformers may use pre training + fine tuning approach on large datasets.

Example Project	Data Type	Feature Type	label Type	Model Type	Algorithm Type	Training, testing & validation
Customer Segmentation	Tabular	Categorical, Continuous	Categorical	Clustering	**k-means:** Minimizes within cluster variance by assigning data points to nearest cluster center **Hierarchical Clustering:** Merges data point iteratively based on similarity	**Hold out validation** on a seperate test set to assess model performance. **Elbow method** used for optimal cluster number selection.

Example Project	Data Type	Feature Type	Label Type	Model Type	Algorithm Type	Training, testing & validation
Price prediction	Tabular	Numerical features	Continuous	Regression	**Linear Regression:** Finds a linear relationship b/w features & targets **XGBoost:** Uses boosted decision trees to capture non linearities in data.	Train-test-split MAE & RMSE used as metrics.