Lecture 2.3

- Error Analysis
  - Train/Test Split, validation set
  - Confusion Matrix
  - Accuracy, Precision, Recall, F-measure, ROC curve,

Train/Test Split in Machine Learning

- **Train-test split** is a machine learning technique that divides a dataset into two subsets: a training set and a testing set
- It's a model validation process that helps assess how well a machine learning model will perform on new data
- Typical Split Ratios
  - 80% for training and 20% for testing
  - 70% for training and 30% for testing
  - 90% for training and 10% for testing (for large dataset)

Validation Set

- The **validation set** is an additional subset of the dataset used to tune the model's hyper-parameters and evaluate its performance during training
- It acts as an intermediary between the training set and the test set
- Purpose of a Validation Set
  - Hyper parameter Tuning
  - Early Stopping
  - Model Selection
- Train/Validation/Test Split
  - Training Set: Used to train the model
  - Validation Set: Used to tune hyper parameters and evaluate the model during training
  - Test Set: Used to assess the final performance on unseen data

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Recall measures the proportion of correctly predicted positive observations out of all actual positives.

$$True\ positive\ rate\ (TPR), recall, sensitivity\ (SEN) = \frac{TP}{TP + FN} = \frac{TP}{P}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision measures the proportion of correctly predicted positive observations out of all predicted positives.

The **F-score** (or F1-score) is a metric that combines precision and recall into a single score, providing a balance between the two. It's especially useful when the data is imbalanced.

$$F - Measure = \frac{2. Precision. Recall}{Precision + Recall}$$

False Positive Rate (FPR) is a measure used in binary classification to quantify how often a model incorrectly predicts a positive outcome for a negative instance

$$False\ positive\ rate\ (type - I\ error) = \frac{FP}{FP + TN}$$

ROC Curve

- An **ROC (Receiver Operating Characteristic)** plot is a graphical representation used to evaluate the performance of a binary classification model
- It illustrates the **trade-off** between the **True Positive Rate (TPR)** and the **False Positive Rate (FPR)** at various threshold settings for a classifier. Here's a breakdown of its meaning and components

$$True\ positive\ rate\ (TPR), recall, sensitivity\ (SEN) = \frac{TP}{TP + FN} = \frac{TP}{P}$$

$$False\ positive\ rate\ (type - I\ error) = \frac{FP}{FP + TN}$$

Components of ROC Curve

- **X-axis**: False Positive Rate (FPR)
- **Y-axis**: True Positive Rate (TPR)
- **Curve**: Plots TPR against FPR for various threshold values
- **Diagonal Line**: Represents a random classifier (no predictive power)
  - The area under this line is 0.5
- **Area Under the Curve (AUC):** The AUC score measures the overall performance of the model
  - An AUC of 1.0 indicates a perfect classifier, while 0.5 indicates a model with no discriminative ability.

Confusion Matrix Generation

Predicted	True
1	1
1	1
1	0
1	1
1	0
1	0
1	1
1	1
0	0
0	0
0	1
1	1
1	0
0	0
0	0
1	1
1	1
1	0
0	1
0	0

	Actually Positive (P) (1)	**Actually Negative (N) (0)**
Predicted Positive (PP) (1)	**8 (TP)**	**5 (FP)**
Predicted Negative (PN) (0)	**2 (FN)**	**5 (TN)**

$$False\ positive\ rate\ (type - I\ error) = \frac{FP}{FP + TN} = \frac{5}{10} = 0.5$$

$$Accuracy = \frac{8+5}{8+5+5+2} = 0.65$$

$$Recall, sensitivity\ (SEN) = \frac{TP}{TP+FN} = \frac{8}{8+2} = 0.8$$

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{8}{13} = 0.62$$

$$F - Measure = \frac{2 \times 0.62 \times 0.8}{0.62 + 0.8} = 0.70$$

ROC Generation

Predicted	True
1	1
1	1
1	0
1	1
1	0
1	0
1	1
1	1
0	0
0	0
0	1
1	1
1	0
0	0
0	0
1	1
1	1
1	0
0	1
0	0